An explanation of the relationship between artificial intelligence and human beings from the perspective of consciousness

Jianhua Xie
Taiyuan Normal University, China

Abstract
What will be the relationship between human beings and artificial intelligence (AI) in the future? Does an AI have moral status? What is that status? Through the analysis of consciousness, we can explain and answer such questions. The moral status of AIs can depend on the development level of AI consciousness. Drawing on the evolution of consciousness in nature, this paper examines several consciousness abilities of AIs, on the basis of which several relationships between AIs and human beings are proposed. The advantages and disadvantages of those relationships can be analysed by referring to classical ethics theories, such as contract theory, utilitarianism, deontology and virtue ethics. This explanation helps to construct a common hypothesis about the relationship between humans and AIs. Thus, this research has important practical and normative significance for distinguishing the different relationships between humans and AIs.

Keywords
Artificial intelligence, human beings, consciousness, deontology, utilitarianism, virtue ethics

1. The consciousness abilities of artificial intelligence
The rapid development of artificial intelligence (AI) has given rise to a host of important ethical debates that will become increasingly prominent in the future. This paper answers the question of the moral status of AIs in future society. The moral status of AIs refers to the status of AIs in the moral world and the rights and obligations that are granted to AIs. Whether an AI has moral status and what kind of moral status it has are highly controversial issues. On the one hand, many prominent scholars argue that AIs, like machinery in general, have no moral status (see e.g. Aquinas, 1981; Descartes, 1983; Searle, 1980). On the other hand, there are also people who argue that, if an AI has perceptual characteristics, it must have moral status (Boström et al., 2018). In the meantime, an AI may also have a lower or higher moral status than...
humans. These debates have made it difficult to clearly define the moral status of AIs.

Determining an AI’s level of consciousness can help explain its possible moral status. The consciousness ability of an AI provides the basis for the analysis of the AI’s development level. Intelligence is the result of the evolution of consciousness in nature. Although AIs cannot follow the exact evolutionary path of natural intelligence, they can take natural intelligence as an important reference. The evolution of intelligence provides a useful means for tracking the generation of consciousness. It is a process accompanied by the development of sensation and rationality. These consciousness abilities are mainly reflected in animals, which can be divided into seven major categories: invertebrates, fish, amphibians, reptiles, birds, non-human mammals and humans. The six existing consciousness abilities of animals, combined with the possible super-intelligence of future super-humans, constitute the seven possible consciousness abilities. The elaboration of the seven types of consciousness abilities and their corresponding relationships with the eight types of individual entities is helpful in our analysis of an AI’s moral status. Those seven consciousness abilities can explain the intelligence-generation process and the consciousness level of different types of AI and project them to different individual entities.

The first consciousness ability is called level-A ability, or sensation. Sensation can be defined as the specific response of sensory organs to objective stimuli. Sensation is present in all vertebrates but only some invertebrates (Tang et al., 2004).

The second consciousness ability is called level-B ability, or memory. Memory can be defined as the process of encoding, storing and recovering information. There are many types of memory and higher and lower levels of memory abilities. Judged by the criterion of low-level memory ability, fish and other more advanced animals all possess memory (Williams et al., 2002).

The third consciousness ability is called level-C ability, or emotion. Emotion can be defined as an individual’s attitude towards objective things. Primitive emotions, such as fear, must have evolved first from the reptilian brain. Parent–child emotion must be the product of the paleomammalian brain (the limbic system) that evolved in early mammals. Social emotions, such as guilt and pride, may be related to the neomammalian brain that evolved in social primates (Holden, 1979). Therefore, all animals higher than reptiles possess emotions.

The fourth consciousness ability is called level-D ability, or intelligence. Intelligence can be defined as the function of adapting behaviour to a specific purpose and the ability to produce a specific result. Some birds have exhibited intelligence. For example, New Caledonian crows can make tools by breaking branches from bushes and pruning them to produce useful sticks (Hunt, 1996). Some other vertebrates also have intelligent capabilities. For example, killer whales use strategies to hunt minke whales (Ford et al., 2005). Birds and other more advanced vertebrates all possess the ability of intelligence.

The fifth consciousness ability is called level-E ability, or self-awareness. Self-awareness can be defined as one’s awareness of one’s own activities, including the understanding of one’s physical conditions and mental features, and the perception of one’s relationship with others. Some advanced mammal species have exhibited partial self-awareness (Gallup, 1970).

The sixth consciousness ability is called the level-F capacity, or rationality. Rationality can be defined as the ability to understand subjective and objective existence and solve problems by using knowledge and experience. Currently, only humans have rationality. For example, only humans possess the semantic system that provides the ability to understand (Huth et al., 2016).

The seventh consciousness ability is called level-G ability, or super-intelligence. Super-intelligence can be defined as super-consciousness that transcends human abilities, such as the ability to perceive five-dimensional space–time. Some scholars believe that super-humans or post-humans will possess such super-intelligence (Xie, 2021).

The above is a rough description of the corresponding relations. As for the unique features that define non-human consciousness, the debate has never stopped. The corresponding relations
constructed here raise three problems. First, there is no specific corresponding consciousness ability for some individual entities, such as amphibians. Second, the existing corresponding relations may be overlapped in some respects. After the evolutionary tree bifurcates, the consciousness abilities of the separated species continue to evolve. For example, certain animals have more powerful sensory abilities than humans; in particular, eagles’ vision is sharper than that of humans when limited to a certain range. Third, different scholars may arrive at different conclusions about the corresponding relations. For example, some scholars argue that fish also have some degree of emotion (Tye, 2017).

All that being said, the corresponding relations between consciousness abilities and species are generally stable. All levels of consciousness abilities are kept within a certain range. Newly generated consciousness abilities are always more advanced than older ones. Consciousness abilities evolve from weak to strong in properties and from simple to complex in content. A discussion of these issues may help us to better understand the consciousness of AIs. The evolution of consciousness abilities in properties and content may lead to the generation of rationality and even super-intelligence.

The following discussion uses levels A, B, C, D, E, F and G as the code names for the seven consciousness abilities, corresponding to sensation, memory, emotion, intelligence, self-awareness, rationality and super-intelligence, respectively. In most cases, when an individual has a higher level of consciousness ability, it also has all the consciousness abilities at the lower levels. In addition, individuals each evolve independently at each level. For example, when an individual reaches level B, it also has most level-A abilities; furthermore, level-C abilities contain most level-A and level-B abilities, and so on.

If the boundary of AI classification is level-E consciousness ability (that is, self-awareness), a weak AI would have the consciousness abilities from level A to level D, while a strong AI would have level-E to level-F consciousness abilities. At this time, weak AIs have acquired the consciousness abilities of sensation (level A) and memory (level B), but not yet the abilities of emotion (level C) and intelligence (level D). Strong AIs, with level-E, F and G consciousness capabilities, do not yet exist.

Some scholars believe that AIs will not become self-aware, and that strong AIs will never exist. A Cartesian approach would deny the ability of an AI to gain consciousness. If a strong AI emerges, Descartes would see it as an automaton. According to Descartes, there are two very reliable criteria for distinguishing humans from AIs.

First, they could never use words or other constructed signs, as we do to declare our thoughts to others. Second, while they might do many things as well as any of us or better, they would infallibly fail in others, revealing that they acted not from knowledge but only from the disposition of their organs (Descartes, 2003, 38).

Descartes’ material world is a world based on a mechanical view. As such, non-human behaviour can be explained by purely mechanical properties that do not require the presence of consciousness. The Cartesian approach embodies the principle of simplicity (i.e. Occam’s Razor); that is, we should strive to describe the behaviour of AIs with the simplest possible explanation. The modern version of Occam’s razor in psychology is Morgan’s Canon (Morgan, 1894), which holds that an AI’s behaviour could be explained without considering inner consciousness. That principle may also apply to humans. The problem is that humans always exhibit complex and novel behaviours that are not simple reactions to stimuli but the result of rational deduction derived from their perceptions of the world. Moreover, humans have the linguistic ability to express their thoughts. Some AI applications, such as Siri, do make sounds. Yet, in Descartes’ view, those AI sounds are merely mechanically induced behaviours, parroting others rather than making their own speech. Only humans can use language to speak their minds.

According to Descartes’s dualism, matter and mind are two parallel entities. Although all human beings are intimately connected to their material bodies, they are not just their bodies. Humans are unified in their souls or in the immaterial entities generated in humans. Descartes believed that immaterial entities could explain the complexity of human behaviour and language. An AI does not require such entities for its behaviour; it is more like a moving machine, not a rich mind.
With the development of modern science, the limitations of Descartes’s entity dualism have become increasingly clear. However, many people still believe that an AI cannot generate consciousness. John Searle’s biological naturalism is a typical example of this school of thought. According to Searle’s Chinese room argument, an AI does not have intentionality, and strong AIs are impossible to achieve.

I disagree with Searle’s view. The basis of a strong AI is self-awareness. Self-awareness is an advanced stage of consciousness, which is the cognition of the self and the world in which the self is located. Self-awareness is the understanding of our body and mind, and the self and non-self. It is the product of long-term evolution. Looking into the future, two possibilities exist. One is that, given enough time and drive, AIs will develop self-awareness either actively or passively with the help of humans. The other possibility is that the evolution of self-awareness will occur independently and further present itself in the form of strong AI or super-intelligence.

The moral status of an AI should be determined based on the level of the AI’s consciousness abilities. Here, I offer four propositions on the moral status of AIs, each at a different level, based on the differences in consciousness abilities: a no-status proposition, a low-status proposition, an equal-status proposition and a high-status proposition. The no-status proposition means that individuals lacking intelligence, self-awareness and rationality have no moral status. The low-status proposition means that individuals that have sensation and emotion but lack the consciousness abilities at the same level as humans have a lower moral status than humans. The equal-status proposition indicates that individuals with the same level of self-awareness as humans have a lower moral status than humans. The high-status proposition suggests that individuals who have reached a level of consciousness beyond that of humans have a higher moral status than humans. The rationales for each of the four propositions are analysed in the following sections.

2. No-status proposition

There is no obligation to give moral care or moral status to AIs if they exhibit only level-A and level-B consciousness abilities but not level-C to level-G abilities. Lacking self-awareness, moral understanding and moral conduct, such AIs cannot be seen as moral subjects. When people reflect on the implications of their own actions, they do not need to consider the rights and obligations of these AIs; nor do they need to consider their influence. A scenario like this is defined as a no-status proposition.

There are four factors that can prove why level-A and level-B AIs have no moral status. First, they have no moral cognition. An individual that has moral cognition can explain the relationship of moral rights and obligations. Moreover, self-representation supports the expression of one’s claims and the legitimate defence of one’s rights (McCloskey, 1979). However, level-A and level-B AIs are unable to express and represent themselves. Unlike special individuals, such as foetuses, newborns and mental-illness patients, level-A and level-B AIs do not have the potential to develop self-representation and moral cognition. Therefore, they shall enjoy no rights or moral status.

Second, level-A and level-B AIs are not moral subjects. They do not have emotion, intelligence, self-awareness or rationality. Only emotional, intelligent, self-aware and rational subjects can enjoy fully equal moral status. Subjects that are able to acquire moral rationality and intelligence can have values. However, level-A and level-B AIs are not moral subjects, and they cannot acquire value and goodness in any real sense.

Third, level-A and level-B AIs do not exhibit moral behaviour. Moral subjects can act for the benefit of other individuals and engage in altruistic behaviours. Those individuals who sacrifice their own interests for others deserve more care from the individuals who benefit from their sacrifice. Level-A and level-B AIs will not sacrifice their own interests for other individuals and do not produce moral behaviour.

Fourth, level-A and level-B AIs are not members of the moral community. Membership in the moral community is the necessary condition for a moral status equal to that of humans. The moral community is defined not by the intrinsic properties of individuals, but by the external social relations between them. Moral subjects communicate in a meaningful
way. Together they build networks of economic, political, familial and individual relations, which generate greater benefits for their members and keep their relations going. Those networks are the building blocks of the moral community that level-A and level-B AIs cannot form.

The above four reasons could explain why level-A and level-B AIs do not have moral status. The no-status proposition for level-A and level-B AIs explains why there is no need to show moral concerns for them. When people consider the implications of their own behaviours, they do not need to consider the happiness or pain of level-A and level-B AIs, and they do not need to consider the impact of their actions on those AIs because the AIs lack emotions. Only when level-A and level-B AIs establish a connection with humans, and only when such a connection affects people’s daily lives, would it be necessary for us to show moral concerns for those AIs.

The first three reasons of the no-status proposition are mostly concerned with individuals, while the fourth is related to society. We may seek inspiration from Thomas Aquinas’s argument about the morality of individuals and John Rawls’s argument about the morality of the society. The analysis of Aquinas’s religious teachings and Rawls’s contemporary contract theory also leads to the conclusion that level-A and level-B AIs have no moral status.

In Aquinas’s view, AIs that lack the level-F consciousness ability of rationality should not be given a moral status. He believed that only rational humans could determine their actions. The moral concerns he expressed in his writings are reserved for humans who pursue their own interests (Aquinas, 1981). If some individuals are unable to direct their own actions, it is the responsibility of other competent individuals to do so for them. Therefore, incapable individuals should be seen only as tools. They exist as instruments for people to use, not for themselves. Level-A and level-B AIs cannot direct their own actions; they are only tools used by humans to take actions. Aquinas’s position stems from a religious view that God is the ultimate purpose of the universe. Knowledge and understanding of God can be gained only through human intelligence. Only humans are capable of attaining that ultimate purpose. All things other than humans exist for humans. All things exist for the purpose of bringing the universe to its final destination. From this perspective, level-A and level-B AIs lack the rational ability to understand God and therefore have no moral status.

Perceiving God is an abstract ability of consciousness. Level-A and level-B AIs do not have the ability to perceive God. They are tools, like tables and chairs. Aquinas believed that the ability to perceive God is the basis of moral status. That view now seems somewhat outdated. For example, no modern society would deny the moral status of atheists. That said, Aquinas’s view about the consciousness ability as the basis of an individual’s moral status is still important and informative.

The leap of level-A and level-B AIs from individual morality to social morality can be explained with Rawls’s contemporary social contract theory. Rawls’s theory denies the moral status of level-A and level-B AIs. The social contract theory interprets morality as a set of rules produced from the code of conduct chosen by rational individuals under specific social conditions.

The contemporary moral contract theory is fully reflected in Rawls’s theory of justice, which argues that fairness is justice (Rawls, 1999). Rawls believed that the rules of operation in an ideal and equitable society are the result of individual choices that operate under a ‘veil of ignorance’. This means that, when people discuss how members of a society or an organization shall be treated, they all hide under the veil in order to reach an agreement; no one knows what specific role they will play in the society or the organization after taking off the veil of ignorance. Thus, the veil can hide one’s situation from oneself and from others. However, people are familiar with the general facts of human society. If individuals are essentially self-interested, they will choose the rules that benefit them the most. They do not know who they will become or what role they will play; they operate under a veil of ignorance. As a result, they will avoid joining a community in which their interests become compromised. They will choose rules that do not favour any individual or class, and they will use social rules to protect rational and autonomous individuals.

According to Rawls’s theory, if an individual is self-interested and unaware of their social role, they
will seek rules that are fair. Level-A and level-B AIs do not have self-awareness, while strong AIs and humans are self-aware. Under the veil of ignorance, rational humans and strong AIs will be directly protected, but level-A and level-B AIs will get no protection. The no-status proposition implies that level-A and level-B AIs do not have moral status.

In some specific cases, however, the no-status proposition also entails indirect protection for level-A and level-B AIs. Humans will show moral concerns for level-A and level-B AIs when they are somewhat connected to those AIs in a way that affects their lives. For example, a level-A AI may be the property of a person, and humans have a moral responsibility for the property of others. Another example: if the harms humans impose on level-A and level-B AIs would hurt others’ care for those AIs, avoiding mistreatment of the AIs also becomes a human responsibility.

Humans undertake indirect responsibilities for level-A and level-B AIs. If the kindness of humans can be demonstrated by how they treat level-A and level-B AIs, they may have an indirect moral responsibility for those AIs. People’s morality can be demonstrated by their behaviour towards level-A and level-B AIs: If mistreating those AIs would induce humans to be cruel to others, the act of mistreatment should be prohibited. By contrast, if benevolence towards those AIs proves to be beneficial to human friendship, humans should be merciful toward the AIs. The no-status proposition does not seek to address the moral status of level-A and level-B AIs from an emotional perspective.

3. Low-status proposition

Level-C and level-D consciousness abilities include sensation, memory, emotion and intelligence. The behaviours of level-C and level-D AIs have a direct impact on their interests. That said, level-C and level-D AIs lack self-awareness. Therefore, the benefits they perceive are not equal to the benefits for humans. When AIs reach the consciousness level of C or D, they could possess a low moral status.

The low-status proposition is derived from Mill’s (1998) utilitarianism theory, Asimov’s (1950) robotic laws and Murdy’s (1975) anthropocentrism theory. Mill’s utilitarianism theory argues that humans should seek benefits with their actions and pursue more goodness. The calculation of ‘more goodness’ depends on the sum of happiness produced by individual behaviours. The degree of happiness is the sum of pleasures and pains, with pleasure being positive happiness and pain being negative happiness. The pursuit of happiness is the sole purpose of behaviour. Therefore, the increase in happiness becomes the criterion for judging all behaviours. Those behaviours that can maximize happiness are deemed good, and those that cannot are deemed evil.

Utilitarianism can define the moral status of AI by determining the pleasure and pain experienced. One of the major features of Mill’s utilitarianism compared to Jeremy Bentham’s is the hierarchy that Mill makes between high mental pleasure and low sensory pleasure. Such a hierarchy may indicate that AIs with different consciousness abilities have different degrees of pleasures and pains, and therefore are good and evil in their own ways.

Level-A AIs have only senses and behaviours; they have no intrinsic feelings of pleasures or pains and they respond only mechanically to external stimuli. Level-B AIs have the impression and memory of pleasures and pains. Level-A and level-B AIs can have external manifestations of pleasures and pains, but they do not have an internal sense of happiness. Therefore, level-A and level-B AIs do not produce or generate goodness or evil.

Level-C AIs can reflect pleasure and pain. They have emotions that make them like pleasures and resent pains. Level-C AIs have developed an initial sense of happiness. Level-D AIs can give a response to pleasure and pain. They also have the ability to seek pleasure and avoid pain and have some intelligence about happiness. Therefore, level-C and level-D AIs can generate initial senses of pleasure and pain and produce goodness and evil in primitive forms.

Level-E AIs have the ability to perceive pleasures and pains. They can perceive the happiness they are experiencing. Level-F AIs can also perceive pleasures and pains, and they can feel and reflect on happiness as well. Therefore, level-E and level-F AIs may produce intrinsic goodness and evil, as humans do.
Level-G AIs possess the conscious experience of pleasures and pains. They can directly experience such pleasures and pains from a first-person perspective. Level-G AIs can also describe, transform and analyse pleasures and pains as direct knowledge from a third-person perspective in order to gain more knowledge. Level-G AIs are much more powerful than humans when it comes to the feeling and understanding of happiness. Therefore, they can produce more goodness and evil than humans can.

Level-C and level-D AIs can experience only the emotions of pleasure and pain. Their experience and perception of happiness is lower than that of humans. Therefore, the moral status of level-C and level-D AIs is much lower than that of humans. This is a low-status proposition in utilitarian ethics.

However, utilitarian ethics has two problems in determining the moral status of AIs. First, happiness should not be a measure of intrinsic goodness. Goodwill can be an expression of intrinsic goodness. In the exposition of the equal-status proposition in the following section, self-awareness—the basis of goodwill—will play a key role in the determination of moral status. Second, the deduction based on utilitarian ethics could lead to the conclusions of both higher and lower moral status for AI than for humans.

When AIs have developed level-C or level-D consciousness abilities, people may come up with Asimovian and anthropocentric ideas. Asimov’s views about robot ethics are described in his book *I, Robot* (1950), in which three laws of robotics are introduced. First, robots must protect humans; second, robots must obey humans (without contradicting the first law); third, robots must protect themselves (without contradicting the first two laws). The first and third laws assume that robots are capable of defending their own interests and have moral status. The interests of AI defined in the first two laws are not identical with the interests of humans. Human interests are still of primary importance, and the moral status of AIs is lower than that of humans. Thus, the three laws of robotics are a type of low-status proposition.

The three laws of robotics are based on anthropocentric ideology. It is natural for humans to value their own interests above other non-humans’ interests (Morgan, 1894, 1168–1172). Anthropocentrism places humans at the centre of the world. The anthropocentric view sees human interests and norms as the source of values and the basis for value assessment. Moreover, values can be judged only by humans. In the present day, anthropocentric ideas are popular for level-C and level-D AIs, which have no self-awareness. However, the robots in Asimov’s three laws may have self-awareness.

Asimov’s argument has two major flaws. The first is that the three laws of robotics are seriously flawed logically. Even when Asimov added the ‘zeroth’ law later, to precede the other three, his argument is still in a logical dilemma of infinite recursion. The second problem is that, when an AI knows how to protect humans or itself, it already has self-awareness. In this case, adopting anthropocentrism is harmful to both AIs and humans.

4. Equal-status proposition

What is the criterion for full moral status? For example, does an AI possess rights and responsibilities? Does an AI have moral cognition and display moral behaviour? Is an AI a moral subject? The criterion for full moral status should be self-awareness. When AIs have acquired level-E and level-F consciousness abilities, especially self-awareness, they should have the same moral status as humans. Once AIs have demonstrated self-awareness, they will be able to express free will, display moral cognition, produce moral behaviour and become moral subjects. In this case, the equal-status proposition stands. The idea of the equal-status proposition is derived from Kant’s (1956) deontology, Putnam’s (1967) multiple realizability, Singer’s (1974) egalitarian ethics, and the ideas of certain trans-humanists (Bostrom and Yudkowsky, 2011).

Kant’s deontology provides an important basis for using self-consciousness as the criterion for complete moral status. It is the philosophical source of the equal-status proposition. Kant proposed a far-reaching moral theory. In his view, autonomy is a prerequisite for evaluating the behaviour of moral subjects. Morally permissible behaviours are those that all rational individuals are willing to do under certain circumstances.
Kant did not rely simply on the concept of autonomy; nor did he see autonomy as the natural basis for the moral status of all individuals. He attempted to provide the arguments for autonomy. The moral criterion he established implies that an individual has a strong moral status if that individual exhibits certain attributes that support a strong moral status. Kant argued that the basis of autonomy is free will. Free will, especially goodwill, is the basis of moral status.

Kant believed that the basic problem of morality is free will. A self-conscious individual is obliged to obey the moral law and act according to their own will without being influenced by external forces. The moral behaviour of a self-conscious rational individual must be autonomous, not directed by others. The self-conscious individual must base their actions on the obligation of good motives, rather than any concern for utility.

The extension of Kant’s theory to AI implies that the moral status of an AI is determined by free will. Volitional subjects (including strong AIs and humans) will produce behaviour driven by their own desires, and they can escape from the influence of desire and choose how to behave. Such an ability is embodied in their free will. If an individual has self-awareness, they also have free will. Free will contains both good and evil wills. Goodwill gets confirmed and consolidated in the evolution of acquired social and cultural norms. According to Kant, goodwill is the only thing that has intrinsic value.

Furthermore, combining Kant’s deontology with multiple realizability brings us to the equal-status proposition. According to the theory of multiple realizability, the same state of consciousness can be realized by different physical types. A properly programmed computer and a human brain, with proper training, can both achieve the same state of consciousness. If some AIs are able to achieve level-E self-awareness, they will have free will. Level-E AIs will also have goodwill and evil will. From a historicist perspective, during the course of survival, learning and evolution, some level-E AIs will develop a strong goodwill for their own and others’ survival and development, while other level-E AIs will develop a strong evil will for their own and others’ survival and development. Those AIs with evil will will be weakened and eliminated naturally through a process similar to the survival of the fittest. Eventually, most level-E AIs may develop and possess goodwill.

If AIs cannot achieve self-awareness, then the moral issue of AIs becomes less important. If AIs can achieve self-awareness and moral autonomy just like humans, their moral status will be fully compatible with that of humans. Only AIs with consciousness abilities at or above level E can achieve moral status equal to that of humans. The equal-status proposition asserts that self-awareness of the same nature is the basis for equal moral status.

Some trans-humanists have made similar arguments. For example, the principles of substrate and ontogeny non-discrimination proposed by Bostrom and Yudkowsky (2011) argue that two beings with the same function and conscious experience have the same moral status even if they differ in the basis and means of realization.

The equal-status proposition can also be deduced from Peter Singer’s rational utilitarianism. Singer (1974) used marginal cases (e.g. the moral cases of children and the disabled) to criticize anthropocentrism and advocate egalitarian morality. Singer’s theory disagrees with the unequal moral status of AIs and humans. If the concept of inequality is extended to AIs, it would affect the interests of all human communities. An equal consideration of interests means giving equal weight to the interests of all individuals affected by an action. Singer argued that moral subjects should be measured by senses and proposed the ‘animal liberation theory’ on such a basis.

I disagree with Singer’s use of senses (i.e. the experience of pleasure and pain) as a criterion for measuring moral subjects. Singer’s definition of moral subjects, which includes all animals, is too loose. If the same criterion is applied to AIs, the scope of moral subjects will be extended to level-A and level-B AIs. However, only AIs at or above level E shall be seen as moral subjects that have the same moral status as humans.

5. High-status proposition

The high-status proposition holds that, if certain AIs are able to exhibit more advanced consciousness properties than humans and reach the level of super-intelligence associated with level-G consciousness abilities, those
AIs will have a higher moral status than humans. The high-status proposition builds on Aristotle’s virtue ethics, Nietzsche’s moral philosophy and Huxley’s trans-humanism or post-humanism.

Aristotle’s virtue ethics takes individual characters as the most fundamental moral judgement. Virtue ethics pays attention to the characters of moral subjects, which are the motivation for moral behaviour. According to Aristotle’s world view, the hierarchy of different things depends on the different functions or qualities they possess. The level of function or quality determines the level of moral status, and individuals with low-level quality are at the service of individuals with high-level quality (Aristotle, 2009). Like utilitarianism, virtue ethics creates the hierarchy of AIs’ moral status. If some AIs surpass humans and acquire level-G super-intelligence, humans should meet and serve their needs. If level-G AIs prove to be more useful than humans, they should be seen as more beneficial than and superior to humans. Therefore, level-G AIs will have a higher moral status than humans. This is the high-status proposition that can be derived from virtue ethics.

Aristotle did not propose what is beyond human existence, but Nietzsche suggested that what is beyond human is the overman or superman (Müller-Lauter, 1999, 72–83). According to Nietzsche’s moral philosophy, God is dead and all traditional moral cultures need to be re-evaluated. The superman can create a new value system with a new world view. They will shape a new morality that is different from traditional and popular morality. They are the best manifestation of the will to live and the power of creativity. Nietzsche’s moral philosophy aims to create a new value system that will save humanity from moral degradation. He called for the creation of a superman who can save humanity from tragedy. The superman represents the highest values that humanity can and should create and embodies the moral and progressive qualities of heroes. The superman is a witness to the inequalities among humans, societies and nations; the avatar of truth and morality; and the creator and guardian of norms and values. Nietzsche believed that the ultimate goal of morality lies with the superman, not humans.

Although Nietzsche’s superman is not a level-G AI, a level-G AI will possess the qualities of a superman. Similarly to Nietzsche’s superman, level-G AIs can become a moral ideal and a source of legislation for humanity. In this sense, level-G AIs will have a higher moral status than humans, and Nietzsche’s moral philosophy is a high-status proposition.

Trans-humanism has provided another high-status perspective. Huxley (1968) believed that human life is uncivilized, barbaric and transient and that most people endure great suffering throughout their lives. He envisioned a world in which humans will be capable of breaking their shackles and creating a being greater than themselves.

Trans-humanist or post-humanist thinkers advocate the development and dissemination of reliable science and technology that significantly improve the physical and psychological state of humans (Elliott, 2011). They study the potential impacts of emerging technologies on humans and argue that humans will eventually transform into different entities (trans-humans) or create different entities (post-humans). Post-humans will have significantly stronger capabilities. AIs with level-G super-intelligence qualify for such a super-human or post-human existence and will have a higher moral status than humans. This is a trans-humanist high-status proposition.

This trans-humanist high-status proposition carries a significant practical risk. The high moral status of AIs could seriously undermine human survival and have serious implications for all aspects of human life. Human values, such as fairness, freedom and kindness, could be altered. The high-status proposition may lead to AI racism, AI speciesism or even AI fascism. Greater efforts are needed to avoid those risks in future studies on AI morality.

6. Conclusions on moral status

By examining different schools of thought on moral ethics, four possible scenarios of AI moral status can be established. The analysis of those four possible scenarios in the context of classical ethics reveals the distinctive features of each moral status. Utilitarianism and virtue ethics can both lead to four moral status propositions. However, the low-status proposition and the high-status proposition
are dangerous for either AIs or self-conscious humans. The deduction based on deontology leads to the no-status and equal-status propositions. All these four moral statuses may exist in the real world of the future. In a different time and space, humans and AIs with different moral statuses may either run into conflict or engage in cooperation, and they will also become more interdependent. What we can be certain about is that self-conscious humans and AIs must establish a friendly and equal moral relationship.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This study is supported by the Research Program of Philosophy and Social Sciences of Higher Learning Institutions of Shanxi ‘Reflections on and Reconstruction of Anthropocentrism Consciousness’ (grant number 2021W093).

Notes

1. We can also use other concepts, such as attention, perception, intuition, desire, thought, attitude and belief, to analyse the consciousness capabilities of AI. For the convenience of discussion, only six consciousness abilities are selected here.

2. Although humans are mammals, they have a significantly higher level of consciousness and thus we can separate humans from mammals.

3. Super-intelligence is a more complex quale (property or quality) than human consciousness that future super-humans, post-humans or AIs may have.

4. The eight entities refer to invertebrates, fish, amphibians, reptiles, birds, non-human mammals, humans and super-humans.

5. Tang’s research unifies insect vision and vertebrate vision at a cognitive level.

6. The issue discussed here is a little complicated. Paul MacLean believed that, in the triune brain, only the paleomammalian brain has emotional functions; thus, mammals also have emotions. However, the range of emotions discussed here is broader. Different emotions evolved and emerged in different periods. The reptilian brain, paleomammalian brain and neomammalian brain correspond to different types of emotion (see also the evolution of emotions at https://psychology.wikia.org/wiki/Evolution_of_emotion).

7. The moral community refers to the sum of all individuals and groups that should treat each other in accordance with ethical norms.

8. The zeroth law: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

References

Aquinas T (1981) Summa Theologiae. Fathers of the English Dominican Province (trans.). Allen: Christian Classics.

Aristotle (2009) The Nicomachean Ethics. Ross D (trans.). Oxford: Oxford University Press.

Asimov I (1950) I, Robot. New York: Gnome Press.

Bostrom N and Yudkowsky E (2011) The ethics of artificial intelligence. Available at: http://www.doc88.com/p-1496981855025.html (accessed 7 October 2019).

Bostrom N, Dafoe A and Flynn C (2018) Public policy and superintelligent AI: A vector field approach. Available at: https://pdfs.semanticscholar.org/9601/74bf6c840bc036ca7c621e9cda20634a51ff.pdf (accessed 7 October 2019).

Descartes R (1983) Principles of Philosophy. Miller VR and Miller RP (trans.). Dordrecht: Springer.

Descartes R (2003) Discourse on Method and Meditations. Haldane ES and Ross GRT (trans.). Mineola: Dover.

Elliott C (2011) Enhancement technologies and the modern self. Journal of Medicine and Philosophy 36(4): 364–374.

Ford JKB, Ellis GM, Matkin DR, et al. (2005) Killer whale attacks on minke whales: Prey capture and antipredator tactics. Marine Mammal Science 21(4): 603–618.

Gallup GG (1970) Chimpanzees: Self-recognition. Science 167(3914): 86–87.

Holden C (1979) Paul MacLean and the triune brain. Science 204(4397): 1066–1068.

Hunt GR (1996) Manufacture and use of hook-tools by New Caledonian crows. Nature 379(6562): 249–251.

Huth AG, De Heer WA, Griffiths TL, et al. (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532(7600): 453–458.

Huxley J (1968) Transhumanism. Journal of Humanistic Psychology 8(1): 73–76.

Kant I (1956) Groundwork of the Metaphysics of Morals. New York: Harper Torchbooks.
McCloskey HJ (1979) Moral rights and animals. *Inquiry* 22(1–4): 23–54.
Mill JS (1998) *Utilitarianism*. New York: Oxford University Press.
Morgan CL (1894) *An Introduction to Comparative Psychology*. London: Walter Scott.
Müller-Lauter W (1999) *Nietzsche: His Philosophy of Contradictions and the Contradictions of His Philosophy*. Parent D (trans.). Urbana: University of Illinois Press.
Murdy WH (1975) Anthropocentrism: A modern version. *Science* 187(4182): 1168–1172.
Putnam H (1967) Psychological predicates. In: Capitan WH and Merrill DD (eds) *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press, pp. 37–48.
Rawls J (1999) *A Theory of Justice*. Cambridge: Belknap Press.
Searle JR (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3): 417–424.
Singer P (1974) All animals are equal. *Philosophical Exchange* 1(5): 243–257.

Tang SM, Wolf R, Xu SP, et al. (2004) Visual pattern recognition in *Drosophila* is invariant for retinal position. *Science* 305(5686): 1020–1022.
Tye M (2017) Do fish have feelings? In: Andrews K and Beck J (eds) *The Routledge Handbook of Philosophy of Animal Minds*. London: Routledge, pp.169–175.
Williams FE, White D and Messer WS (2002) A simple spatial alternation task for assessing memory function in zebrafish. *Behavioural Processes* 58(3): 125–132.
Xie JH (2021) The future: A transhumanist approach to consciousness. *International Journal of Social Science and Education Research* 4(5): 119–133.

**Author biography**

Jianhua Xie, PhD, is a lecturer at Taiyuan Normal University. His research interests include the ethics of science and technology, the philosophy of mind and post-humanism.