This paper was presented at a seminar on "AI and the Human Person: Chinese and Western Perspectives," held at Santa Clara University, April 4-5, 2019, sponsored by the China Forum for Civilizational Dialogue (Georgetown University and La Civilta Cattolica), Tech and the Human Spirit Initiative (Santa Clara University), and the Pontifical Council for Culture

AI and the Confucian Conception of the Human Person
Some Preliminary Reflections

The understanding of autonomy and rationality central to the development of AI, and also assumed by cultures influenced by the European Enlightenment to be central to what it means to be human, does not have the same importance in Confucian understanding of the human person. For Confucians, what is most central to our humanity is inter-personal relations. One can be human only in relation to other human beings. Living well is a process of cultivating one's person through one's relationship to family members, friends, neighbours, colleagues, fellow citizens, any other human being one comes into contact with, and conducting oneself virtuously in those relationships, from the most intimate to the most fleeting. While not denying the conveniences and new opportunities of all kinds brought about by AI, from a Confucian perspective, AI magnifies the bias towards the Enlightenment view of autonomous rational human agent and thereby carries the risk of distorting human development. As we increasingly worry about being outstripped in intelligence by AI, and strive to integrate AI with human processes to benefit rather than threaten human welfare, with both intelligence and welfare understood in ways that emphasize the value of Enlightenment autonomy and rationality, we increase the peril of neglecting other aspects of our humanity.

The latest development of AI has moved beyond narrow intelligence limited to specialized tasks. Personal robots, such as Softbank's Pepper, described as "the first social humanoid robot able to recognize faces and basic human emotions" is "optimized for human interaction and is able to engage with people through conversation and his touch screen." According to the manufacturer, "Over 2000 companies around the world have adopted Pepper as an assistant to welcome, inform and guide visitors in an innovative way." [1] This is not just another technology to address a labor shortage problem comparable to automation in manufacturing processes. AI-human interactions are replacing inter-human interactions in societies experiencing labor shortage in some sectors such as education and elder care. Could one still become an exemplary person and live a virtuous life on Confucian terms in a world where AI-human relations replace inter-human relations in more and more activities?

Hitherto, the mundane activities of everyday life bring us into contact with a variety of other human agents, from supermarket cashiers, taxi drivers, waiters, to teachers, nurses, and doctors. We may think that, if the tasks of these various roles could be carried out just as effectively and efficiently, perhaps more so, by robots, then the substitution is no more problematic than having any task performed by one qualified person instead of another. Insofar as the goals can be clearly specified, and determinate paths towards their achievements under varied determinable conditions could be constructed in a system, AI can replace humans, and even out-perform them. However, from a Confucian perspective, while inter-human interactions often have functional aspects, what makes them critical to a person's emotional, intellectual, and moral growth, in other words critical to sustaining her humanity, is not the achievement of those functions but because, beyond the

[1] Softbank Robotics web-page. https://www.softbankrobotics.com/emea/en/pepper. Accessed 1 Mar 2019.

functional goals of any interaction between people in different roles – driver and passenger, doctor and patient, teacher and student – those involved are also interacting with one another as persons, each with his or her particular personality. A transaction, in which one party provides a specific service to another by meeting determinate functional goals, becomes human interaction only when both parties recognize and respond to the unique personality and shared humanity in each other.

In education, technology has provided better ways of measuring student abilities, better and more methods of delivering content to students starting with different abilities, facilitating self-learning and learning at each student's own pace, tracking students' progress, and enhancing students' engagement with content, teachers, and peers, to name but a few examples. With the advance in AI, can we dispense with human instructors altogether? A good teacher does not only deliver content, or rely completely on external measures of a student's abilities and needs, compare them with specified learning objectives to map out the path for the student to get from point A to point B, and help the student accomplish that journey with appropriate instructional tools. Those are the functional goals that define the teacher's responsibilities. Students often forget what they learn in school, whether or not those lessons were taught well and remain relevant. Teachers who merely fulfil the functional goals of education leave little impression as educators; the teachers whom students remember as having made a difference to their lives are those who have made a personal connection and contributed something to their transformation as a person, beyond the specific knowledge gained and skills acquired.

A good teacher gets to know her students personally, understanding the contexts for their measured abilities, their varied interests, attitudes to learning, the difficulties and challenges each is likely to encounter. Such contexts enable her to respond to students' words and actions beyond their superficial meanings, and understand what is left unsaid or lies beneath an action for a particular student in each instance of engagement. Besides the specified learning objectives of any course of study, education is a holistic endeavor of nurturing growth, and it is better achieved when one understands students through interacting with them as fellow humans with unique personalities and experiences. Such interactions go beyond linguistic and logic driven communication and require understanding of each student as a unique person in her own particular circumstances, only then could the educative process have the potential to transform not only the student but also the teacher. This transformative potential, while most evident in teacher-student relationships, is present in all human relationships. Do we understand such processes well enough to design systems that could imitate such processes? Is it even possible for AI to imitate this dimension of human interaction?

Given the importance of the virtue of filial piety in Confucianism, the use of AI in elder care is of great interest. Societies, such as Japan, with rapidly aging population and labor shortage have turned to AI for solutions. Electronic care givers are being used to monitor the needs of the elderly as well as provide mental engagement. Beyond the practical tasks of ensuring their physical health and comfort, and assisting the elderly in daily tasks which have become difficult or impossible for them, robots that can converse, play games, and dance with the elderly are now available on the market. Any new technology to improve the quality and lower the costs of care for the elderly is surely welcome. And from the perspective of how lonely many elderly have become in the fast paced societies today where younger family members have demanding jobs or live too far away to offer day to day companionship and care, even when they would like to do so, the value of such AI substitutes cannot be dismissed.

However, Confucians would question whether we are too quick to accept the inevitability of a future where elderly parents rely on robots rather than their children for care and companionship,

and not devoting comparable resources to explore the possibilities of new socioeconomic arrangements in order to sustain the practice of filial care for parents by children themselves. A son's or a daughter's filial responsibility is not just about satisfying the needs and desires of the parents, its fulfilment is an important part of what it means to be a son and daughter, to be a human person. Paying for the best technology, or for that matter paying for other people, to care for one's parents is no more satisfactory in discharging that responsibility than paying for one's offspring's material upkeep alone is adequate parenting. It is of course not realistic to expect every child to be able take care of his or her parent single-handedly, assistance from others is always needed. As another means of improving what we can do in caring for the elderly, AI is welcome; but it should not become accepted as a convenient substitute for the personal caring children owe their aging parents.

Some might even think that replacing inter-human interactions with AI-human interactions in care industries and even in homes can provide superior care to the extent that one can ensure that a robot has all the made-to-order qualities of an ideal care giver or companion, such as inexhaustible patience, cheerfulness, humor, sympathy, and so on, compared to interacting with another imperfect human being who cannot avoid at least occasionally falling prey to frustration, bad moods, or being distracted by his or her own personal troubles. However, this treats the relationship as purely instrumental and overlooks the human dimension. It misunderstands the nature of our human need for care and companionship. An important aspect of human relationships is reciprocity. This may not seem to apply to relationships in which one party is dependent on the other for care. However, this one-way dependence and its corresponding inequality do not preclude the important reciprocity rooted in shared humanity. Reciprocity in this sense does not require equality of abilities or equal exchange. To be truly human, it is not enough to have interaction with another who/which caters completely to one's needs and desires. One must in turn respond to the other as a unique human person, not just as the means to satisfying one's needs and desires. The means can be replaced whether by another person or perhaps by a machine, but the person cannot be replaced. Recognizing and responding to the humanity of the person who cares for oneself is to recognize and accept her imperfections, and to appreciate her care despite, and even because of, those imperfections.

It is true that not every one of our encounters with another human being is a human relationship with this kind of reciprocity, in which we relate to the other as a unique human person. Being imperfect, every one of us at least some of time, and perhaps even most of the time, treats others as means to satisfy our ends. When I go to the information counter to seek help, all I want is accurate information, and if a robot could provide that information, and furthermore is never rude or unpleasant, there is no reason to prefer a human person to AI from the perspective of achieving my goal. AI-human interactions are superior to inter-human interactions in being free from human errors; but they are also incapable of initiating and responding to unexpected actions that could enrich human experience. On one occasion, the person behind the information counter noticed that I did not look well, and besides answering my question, enquired after my well-being and whether I might need other assistance. A robot might be able to add a general enquiry about one's well-being, like a polite greeting, but would it be able to make that *personal connection* which comes from attentiveness to a particular person beyond that person's expressed needs or desires? Even though most of time we may approach others with some specific purpose in mind, we can receive more than achievement of that purpose from the other when the interaction becomes more than a strictly goal-oriented transaction because parties involved connect with each other beyond that specific purpose by recognizing and responding to each other as unique particular persons. What would it be like to live a life so goal-oriented that it dispenses with all human interactions? Can AI-human

interaction advance beyond goal-oriented transactions to interactions between unique particular persons?

The Pepper robot is advertised as able to engage humans through conversation, but what kind of conversation? Can a robot give an appropriate rather than a correct response when conversing? Correct answers are available when one is required to give information or guide another in a specific task, but meaningful conversations between human beings are constituted by appropriate responses that are as much about the person who asks the question or makes the initial comment as it is about the one who answered or responded. An elderly person's need for companionship is not just the need to have someone listen to her. She needs someone in whom she could take an interest, someone who could share with her that person's daily experience in the world in which she is no longer as active, someone who has different perspectives, and life experience. Her interest is interest in another human being, who in turn can be interested in her as a fellow human being. The elderly desire conversations in which each could share his or her own life experience with another person who not only shows attentive interest, who is interested in her as a unique person, but who hopefully could also benefit as another imperfect human being who must make their way through a world that is not created to cater to his or her every need and desire.

The human dimension of interpersonal relations is indispensable in education and relations of care. It is not only the very young or the elderly among us who need care and companionship. However strong and independent a person, as a human being, we need such relations with other human beings, and our continued emotional, intellectual, and moral growth, even more than our physical growth from child to adult, depends on a variety of relationships with other human beings, in which each party contributes to the humanity of the other and vice versa. Unless AI could imitate humans to such an extent of having comparable life experiences and perspectives, and participating in the mutual constitution and reciprocal enrichment of humanity in relationships, it cannot provide adequate substitutes in tasks that by their very nature have and need this human dimension. The more AI-human interactions replace inter-human interactions in our daily life, the less opportunities we have of the experiencing and participating in the mutual growth of our humanity. However desirable the functional assistance AI could provide, failing to recognize its limitations imperils our very humanity.

A central tenant of Confucius' teaching is "Do not impose on others what you yourself do not want." (*Analects* 15.24) This requires putting oneself in another's position, imagining how one would feel in that situation if one were the other person. Hence, empathy is critical in human relationships and seems to be embedded in what we share as a biological species with common physical, emotional, and intellectual characteristics, capacities, limitations, and vulnerabilities. Empathy also arises from common experience and culture within disparate groups of human beings. What we expect from another human being often is not only the fulfilment of functional goals we desire, sometimes what we desire more is *understanding*. Not factual knowledge of our situation, or rigorous reasoning of causes and implications that could lead to solutions if the situation is problematic. What we desire is understanding in the sense of appreciating what we are experiencing, our very human joy or suffering, fear or anxiety.

A robot that can "recognize basic emotions" is a big step the direction of AI-human interaction approximating human interaction. How close is this to AI that can empathize with humans? The use of AI in areas demanding interaction with human beings that imitates inter-human interaction has increased attention paid to the integration of emotion into systems that think and act autonomously like humans. Such affective engineering develops systems that can recognize and respond to human emotions by simulating the appearance of emotions, but has not progressed to

the stage of instantiating human emotions themselves in the systems, assuming that the latter is possible without the same, or even any, biological substrate. Further progress would require significantly advancing our understanding of emotions, which have been peripheral and overlooked not only by AI researchers, but also cognitive scientists, philosophers of mind and many others. There is little agreement, *inter alia*, on what is an emotion, the appropriate taxonomy of all the things we call emotions, how they are related to consciousness, rationality, motivation, desires, intentionality, and the roles they play in actions. Due to the Enlightenment legacy, current efforts to integrate emotions into AI tend to privilege the cognitive in approaching emotions, even reduce the latter to the former, and prioritize reason-based intelligence over emotions in both their roles and their values in decision making and actions.

Compared to the prevalent thinking in AI, reason and cognition does not have the same priority in Confucianism. On the contrary, the primacy of emotions in Confucian understanding of the human person and Confucian ethics leads to an emphasis on the cultivation of emotions in education. Mencius, second only to Confucius in the Confucian tradition, distinguished human beings from animals by their possession at birth of the "four sprouts" of humaneness (*ren* 仁), appropriateness (*yi* 义), ritual (*li* 礼), and wisdom (*zhi* 知) (*Mencius* 2A6). These are respectively the heart-mind of compassion, the heart-mind of shame, the heart-mind of courtesy and modesty, and the heart-mind of right and wrong. The translation of "heart-mind" has been adopted for *xin* (心), a term which in the Chinese language refers to the heart as a physical organ, but also as the seat of thought and emotion. The term in that key passage has also been translated as "feelings" to signify the absence of thought in the four sprouts, which are presented as immediate, spontaneous response to specific situations. We cultivate these sprouts when we learn how to relate to other people, and the success of such cultivation would result in the excellences of an exemplary person. Such cultivation does not involve only mental efforts, but engages the entire person; the body is not separate from the mind and is as important, perhaps more important, than the mind in Confucian personal cultivation. This is clearly evident in the importance of ritual education and conduct in the Confucian tradition. Based on the Confucian understanding of emotions and its role in cultivating and conducting oneself as a human person, adequate AI-human interaction must go beyond processing linguistic expressions of emotions, treating them as symbolic structures that can be manipulated to yield some response that imitate human emotions. Both the perception and simulation of human emotions in AI must be more holistic if they are to meet Confucian standards.

Raising the above concerns from a Confucian perspective is not intended to reject AI outright, as Confucianism has no doctrinal resistance to modern technology, and has a long history of adapting to the civilizational transformations brought about by successive waves of scientific and technological advances. The intention is to suggest issues that further developments of AI need to take into account to avoid some pitfalls. Not everyone would agree with the Confucian perspective, but critical reflection and open discussions of these issues are vital if we are to make AI more human, capable of enhancing rather than threatening our humanity. The danger of not doing so is to allow the often hyped-up promotion of AI to erode our understanding of humanity, so that rather than clarifying what AI still lacks that makes it not human in significant ways, and perhaps makes it impossible for AI ever to be human, we allow what AI could imitate and aims to imitate to limit our understanding of what it means to be human.

Sor-hoon Tan
Professor of Philosophy
Singapore Management University