International Psychogeriatrics (2020), 32:8, 909–911 © International Psychogeriatric Association 2020
doi:10.1017/S1041610220001684

COMMENTARY

How to make Artificial Wisdom possible

In advocating for the development of Artificial Wisdom (AW), Jeste *et al.* (2020) lay a foundation in understanding intelligence and wisdom and their importance to humanity. They outline the development of recent successes in the field of Artificial Intelligence (AI) and our scientific understanding of human intelligence and wisdom. With this as grounding, the discussion of the construct of human wisdom is taken as the basis for proposing governing principles for the development of AW. However, while there may be substantial benefit in having computers with AW, it is important to ask whether it is scientifically feasible to develop true AW that mimics human wisdom.

In the midst of an airborne viral pandemic, wearing a mask can be smart. In the midst of a health crisis in which people are dying, lying about mortality statistics is not smart, especially given that in a democracy the truth will come out eventually. Although neither of these situations seems on its face to be about wisdom, could AI tell the difference between a smart and a stupid choice in these cases? Although the situations are presented as if there are clear smart choices, is it true that one choice would be always be judged by everyone to be smart? In the present world context, some political advisors might say that given the probabilities of infection (low) and the need to keep morale positive (high) and possible negative implications of the appearance of someone wearing a mask, perhaps it would be smart for leaders to not wear masks, even if this models behavior is not smart. Furthermore, the same political advisors might say that in the short term, the political benefit of good news about a crisis and reducing bad news such as increased mortality by shading statistics outweighs the principle of truth and transparency. Thus, society's values for public goods such as health and truth can be in conflict with individual values for political goods. When values are in conflict, leaders often make choices that are clever or smart about their own needs, but are often not wise. Could an AI system make the smart choice, or the clever choice? Is it even possible for a computer to determine what a wise choice would be?

Although Jeste *et al.* (2020) state that the ultimate goal of AI is to serve humanity, while that is an aspirational goal for some such as the Future of Life Institute (https://futureoflife.org/ai-open-letter/), and the concerns about the potential societal threats of AI (e.g., https://futureoflife.org/open-letter-autonomous-weapons/), research on AI does not have a single overarching goal. While there are engineers seeking to use AI to address problems facing humanity such as expert systems to improve medical diagnosis like Caduceus (e.g., Banks, 1986) and other beneficial systems, there are also engineers developing systems, even outside of the weaponizing of AI that concerns the scientists represented affiliated with the Future of Life Institute, systems that can threaten jobs or generally present competition for humanity (see https://www.newyorker.com/tech/annals-of-technology/why-we-should-think-about-the-threat-of-artificial-intelligence). Still other researchers take the development of AI as a scientific approach to understand human cognition (see Forbus, 2010), as a form of Cognitive Science without any interest in serving humanity but just as a method of scientific research on psychological processes.

In the early days of AI research, the first goals for researchers were to produce programs that could emulate smart human behavior such as producing intelligent answers for IQ test-like problems or playing games like chess or checkers at human levels of excellence (see Nilsson, 2010). However, over time, researchers realized that it is possible to reach human levels of performance on narrowly specific tasks such as playing checkers or solving word problems, but these solutions can be achieved without a deep understanding of human cognition. They are really only brittle surface simulations of behavior. Even more complicated examples such as Colby's (1973) Parry model which was intended to simulate human paranoia, only used a simple set of rules that operated on patterns of words in sentences. While sufficient to pass a modified version of the Turing test (Turing, 1950), Parry only mimicked the symptoms of paranoia without a clear theory of paranoid ideation; there was no real model of the internal thoughts and feelings of paranoia. This approach was parodied by Weizenbaum (1974) by showing that mimicking the appearance behavior does not advance a scientific understanding of human psychology (also see Colby, 1981, and commentaries). This points up a fundamental problem with the use of a Turing test advocated by Jeste *et al.* for assessing fundamental psychological processes like intelligence or wisdom: The Turing test only assesses the putative output expression of a process

(Church-Turing equivalence; e.g., Soare, 1996), but such expression can be functionally simulated without the underlying aspects of the process that are most critical. This means that outside such a test, there can be deviations in performance, especially when the demands of a situation become more substantial challenges.

From simple surface emulations of behavior of complex psychological processes such as paranoia, it was clear that AI needed to model the mechanisms underlying general cognitive abilities like analogical reasoning or language comprehension rather than the manifest behavior. However even as AI has improved in performance in game playing, beating experts at checkers then chess and most recently go – considered a long-standing challenge for AI (Singh *et al.*, 2017) – and improved in problem-solving, pattern recognition, and translation, AI performance is still at the level of human abilities for performance that depends on aspects of human life that cannot be described in information theoretic terms such as emotion. In part, this may be due to the limited scientific understanding of the basic processes that underlie natural human intelligence is (e.g., see Searle, 1980, and commentaries). If the accurate modelling of human intelligence remains a challenge for AI, will wisdom represent a greater challenge? Clearly Jeste *et al.* believe AW is a solvable problem.

However, as Jeste *et al.* point out, human intelligence does not have a single agreed upon definition, although as they note, there are a number of standardized IQ tests that presumably assume such definitions. It is important to note that Binet, the father of the modern IQ test, described the faculty of intelligence as judgment or good sense (Binet and Simon, 1916). Thus, it is problematic that standardized IQ tests as identified by Jeste *et al.* are not tests of judgment or good sense. The practice of measuring intelligence does not accord with the much richer and more complex conception that Binet about intelligence. This reflects compromises that are made to operationalize complex human psychology.

Furthermore, prudential judgment or good sense, the synonyms offered by Binet for intelligence might be associated with wisdom, at least per Aristotle in the Nichomachean Ethics Book VI. When considering the recent intelligent behaviors demonstrated by AI referred to by Jeste *et al.*, these behaviors are probably not what most people would think of as demonstrating good sense. Rather these are closer to a more limited notion of problem-solving, even when impressive deep learning or inductive reasoning is involved.

It is interesting to note that some researchers have identified a place for wisdom in computer science. The DIKW pyramid (https://en.wikipedia.org/wiki/DIKW_pyramid) represents a relationship among Data, Information, Knowledge, and Wisdom, which bears an interesting metaphoric relationship with human wisdom. Data refer to the numbers or the observations that might be made. Information might be thought of as the context within which the observations are made. Knowledge might be conceived of as the meaning of the observations or the functional use of the numbers and wisdom could be described as the use of values to guide knowledge use. This notion of values either as motivation or goals is interesting because the idea of human wisdom being grounded in values is important (Tiberius, 2008). This idea that values should shape the use of knowledge appears to be common to both the psychological science (Grossmann *et al.*, 2020) and computer science notions of wisdom. In fact, as with psychological science and philosophy, Jeste *et al.* have identified several moral virtues as guiding values which they call principles for the development of AW.

In some views (see Schwartz and Sharpe, 2006), wisdom is conceived of as a master virtue that works to organize or mediate other virtues such as empathy. By contrast, Jeste *et al.* identify moral virtues as important guiding principles for AI, suggesting that these are necessary preconditions for wisdom. These moral virtues are common across a number of theories of wisdom in humans (e.g., Grossmann *et al.*, 2020): (1) the importance of reflection and perspective taking, (2) empathy and compassion, and (3) emotional understanding and emotional self-regulation. As guiding principles for the development of AW, as opposed to virtues managed by wisdom, Jeste *et al.* are asserting these as foundational abilities necessary for the manifestation of wisdom. However, if viewed as necessary foundations for wisdom, these virtues are in domains that are most difficult for computer science to advance. In this respect, Jeste *et al.* importantly reflect this challenge and suggest the importance of human–machine interaction in developing AW as well as the critical point that AW needs to be adaptive based on experience. Rather than conceiving the wise computer, they suggest that whatever starting point AW has, which could be remote from what would constitute human wisdom, feedback from experience and specific interactions with humans could lead to wiser subsequent performance. This seems like a fundamentally important concession and, as they indicate absent human capacities such as consciousness and emotion, it may be better to think of AW as emerging from a partnership with humans rather than a standalone system of wisdom.

The traditional view of AI is that it is typically conceived of as operating as an autonomous intelligent agent. However, given the challenge of giving a computer moral virtues such as the ability to reflect,

to take perspective, to feel empathy and compassion, and to understand deeply emotional responses, perhaps this individualistic model of an independent wise agent in AW is not the right perspective. As suggested by Jeste *et al.*, AW needs to emerge in partnership with humans, perhaps as advisor or consultant to complement human partners' understanding and reasoning. In this way, it may be that AW is not actually manifested by the software but instead by the combination of human and software avatar. This suggests that the idea of AW may be more akin to "institutional wisdom" (see Nusbaum, 2019), in which institution's policies lead its constituents to make wiser decisions (cf. Thaler and Sunstein, 2008) but there may be no individual in the institution that is individually wise. In this view of institutional wisdom as emergent cognition, the whole of the governing body of the institution is wiser than any person in the group. This notion of AW is that even if a person is not wise, a system composed of person plus computer can be wiser than either alone. This is a powerful new way of thinking about the development of a unique form of expert system that benefit humanity wisely.

Howard C. Nusbaum

Center for Practical Wisdom, The University of Chicago, Chicago, IL, USA
Email: hcn1@uchicago.edu

References

Banks, G. (1986). Artificial intelligence in medical diagnosis: the INTERNIST/CADUCEUS approach. *Critical Reviews in Medical Informatics*, 1(1), 23–54.

Binet, A. and Simon, T. (1916). *The Development of Intelligence in Children* (pp. 42–43). Baltimore, MD: Williams & Wilkins.

Colby, K.M. (1973). Simulation of belief systems. In: R. Schank and K. Colby (Eds.), *Computer Models of Thought and Language*. San Francisco, CA: W. H. Freeman.

Colby, K.M. (1981). Modeling a paranoid mind. *Behavioral and Brain Sciences*, 4, 515–534.

Grossmann, I. *et al.* (2020). The science of wisdom in a polarized world: knowns and unknowns. *Psychological Inquiry*, 31(2), 103133. doi: 10.1080/1047840X.2020.1750917.

Forbus, K.D. (2010). AI and cognitive science: the past and next 30 years. *Topics in Cognitive Science*, 2, 345–356. doi: 10.1111/j.1756-8765.2010.01083.x.

Jeste, D., Graham, S., Nguyen, T., Depp, C., Lee, E. and Kim, H.-C. (2020). Beyond artificial intelligence: exploring artificial wisdom. *International Pyschogeriatrics*, 32, 993–1001. doi: 10.1017/S1041610220000927.

Nilsson, N. (2010). *The Quest for Artificial Intelligence*. Cambridge: Cambridge University Press.

Nusbaum, H. C. (2019). The breakdown of civic virtues and the problem of hate speech: Is there wisdom in freedom of speech? In: R. Sternberg, H.C. Nusbaum and J. Gluck (Eds.), *Applying Wisdom to Contemporary World Problems* (pp. 111–142). London: Palgrave Macmillan.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, 3, 417–457.

Schwartz, B. and Sharpe, K.E. (2006). Practical wisdom: Aristotle meets positive psychology. *Journal of Happiness Studies*, 7(3), 377–395.

Singh, S., Okun, A. and Jackson, A. (2017). Learning to play Go from scratch. *Nature*, 550, 336–337. doi: 10.1038/550336a.

Soare, R. (1996). Computability and recursion. *Bulletin of Symbolic Logic*, 2(3), 284–321. doi: 10.2307/420992.

Tiberius, V. (2008). Introduction. *The Reflective Life: Living Wisely with Our Limits* (ch. 1, pp. 3–18). Oxford, UK: Oxford University Press.

Thaler, R. H. and Sunstein, C. (2008). *Nudge*. New Haven, CT: Yale University Press.

Turing, A. (1950). Computing machinery and intelligence. In: E.A. Feigenbaum and J. Feldman (Eds.), *Computers and Thought*. New York: McGraw-Hill.

Weizenbaum, J. (1974). Automating psychotherapy. *Communications of the Association for Computing Machinery*, 17(7), 425.