**Quantum of Wisdom**

Colin Allen & Brett Karlan

Abstract: Practical quantum computing devices and their applications to AI in particular are presently mostly speculative. Nevertheless, questions about whether this future technology, if achieved, presents any special ethical issues are beginning to take shape. As with any novel technology, one can be reasonably confident that the challenges presented by "quantum AI" will be a mixture of something new and something old. Other commentators (Sevilla & Moreno 2019), have emphasized continuity, arguing that quantum computing does not substantially affect approaches to value alignment methods for AI, although they allow that further questions arise concerning governance and verification of quantum AI applications. In this brief paper, we turn our attention to the problem of identifying as-yet-unknown discontinuities that might result from quantum AI applications. Wise development, introduction, and use of any new technology depends on successfully anticipating new modes of failure for that technology. This requires rigorous efforts to break systems in protected sandboxes, and it must be conducted at all stages of technology design, development, and deployment. Such testing must also be informed by technical expertise but cannot be left solely to experts in the technology because of the history of failures to predict how non-experts will use or adapt to new technologies. This interplay between experts and non-experts may be particularly acute for quantum AI because quantum mechanics is notoriously difficult to understand. (As Richard Feynman quipped, "Anyone who claims to understand quantum mechanics is either lying or crazy.") We will discuss the extent to which the difficulties in understanding the physics underlying quantum computing challenges attempts to anticipate new failure modes that might be introduced in AI applications intended for unsupervised operation in the public sphere.

With IBM, Microsoft, Google, and Amazon all working to bring quantum computing to their cloud computing platforms, the projected benefits of quantum computing may seem poised to hit the mainstream. Yet numerous technical challenges remain for the development of quantum hardware and the algorithms to run on quantum computers. Physical hardware for quantum computing requires special methods for preserving the coherence of quantum states that are easily disturbed. The number of quantum bits (“qubits”) in quantum computers has grown (on IBM’s platform, for example, from 5 qubits in 2016 to 16 qubits at the end of 2020), and programmers have been given greater control over the topology within which the qubits interact. Yet this is far short of the kind of exponential growth in computing power described by “Moore’s Law”, which has characterized computer engineering in the decades since microchips were first developed. In recent years, this growth has shown signs of slowing as physical limits to miniaturization of transistors have been approached. This has led some commentators to declare the end of Moore’s Law (e.g. Theis & Wong 2017). Others have pointed to quantum computing as the next technological development that will keep it going. However, both hardware design and the development of quantum algorithms to exploit the inherent parallelism of quantum states face tricky issues. Practical quantum computing thus remains speculative, and applications to artificial intelligence (“quantum AI”) are even more speculative.

The authors of this short article are philosophers of cognitive science whose interest in AI (including machine learning) spans the use of AI for cognitive and scientific modeling, and the ethical impacts of deploying AI in various online and robotic applications. We are currently pursuing a project concerning how the mismatch between AI capacities and human understanding of those capacities presents barriers to wise use of the technology. Our particular focus in this article is on whether quantum computing presents any special issues for the ethics of AI. We are not concerned here with speculation about whether quantum effects are integral or essential to human intelligence or consciousness.

Despite the speculative nature of quantum AI, questions about whether this future technology presents any special ethical issues are beginning to take shape. As with any novel technology, one can be reasonably confident that the challenges presented by quantum AI will be a mixture of something new and something old. What little literature exists on this topic so far emphasizes continuity. For example, Sevilla & Moreno (2019) argue that quantum computing does not substantially affect methods for achieving value alignment between AI and humans, although they allow that further questions do arise concerning governance and verification of quantum AI applications.

In this brief paper, we turn our attention to the problem of identifying as-yet-unknown discontinuities that might result from quantum AI applications. Quantum mechanics is notoriously difficult to understand. (As Richard Feynman quipped, "Anyone who claims to understand quantum mechanics is either lying or crazy.") Insofar as quantum computing rests on some of the mysterious aspects of quantum mechanics, and insofar as the various aspects of intelligence, whether natural or artificial, remains only dimly understood, quantum AI's position at the nexus of these might be thought to present novel challenges to its ethical use.

To understand the possible discontinuities and continuities of ethical questions in quantum AI, however, we first need a framework for thinking about ethical questions in AI in general. We are currently engaged in a project to frame issues of AI-human interaction in terms of *practical wisdom*. Our conception of practical wisdom with respect to technological artifacts has two main dimensions: broad and deep *knowledge* of the system's operating characteristics, on the one hand, and metacognitive awareness about the *limits* of that knowledge, on the other (Karlan & Allen, in prep.). We argue that people engaged in various and often multiple roles -- as designers, engineers, programmers, managers, salespeople, customers, and end users, as well as regulators and the public at large -- need different combinations of knowledge and support for developing appropriate understanding and metacognition.

Numerous examples from AI already demonstrate the necessity of approaching problems with both a wide breadth of knowledge about the relevant conditions and metacognitive awareness of the shortcomings of that knowledge. State-of-the-art neural networks, for example, are vulnerable to adversarial attack, often via manipulations that are imperceptible to human perception (Szegedy et al. 2014). When using a computer vision algorithm to direct a self-driving car, how can programmers and users know when the car is likely to be fooled by a naturally occurring "edge case" that has not been seen during training or an adversarial attack

introduced by another agent, or to make inaccurate judgments due to biases in its training data? Answering this question is complicated, but of paramount importance is knowledge of both the workings of the car and the exact instances in which blindspots and biases are likely to be consequential, or in which adversarial manipulation is likely to be encountered. (For more examples, see Karlan & Allen, in prep.)

In the context of quantum computing, identifying the shortcomings of our knowledge is made significantly more difficult by the opacity inherent to the quantum system itself. The potential ethical pitfalls are nonetheless important to understand: if a self-driving car operated by a quantum computing architecture crashes and causes loss of human life, it might be physically impossible to recover information from the system without altering the system in unrecoverable ways. The ethical implications of this kind of opacity are multiple and worrying, and we reflect on them in detail below.

A practical-wisdom perspective on the use of AI focuses stakeholders at all levels on anticipating new modes of failure in the technology. One can know many things about how an AI product will work without knowing the first thing about how, and in what ways, it will fail, as in the adversarial attack example above. Practical wisdom requires one to be aware of the limitations of an AI as well as its strengths. It also requires being aware of one's own limits in understanding the limitations of the technology. To increase the practical wisdom of humans working with AI will require building sandboxes where stakeholders can evaluate the functioning of an AI within safe parameters, outside of real-world stakes and consequences of failure, and other spaces where people can brainstorm about potential problems and solutions without fear of repercussions. These sandboxes and spaces for discussion among multiple participants must be developed for all stages of product design, implementation, and deployment to ensure that wise design permeates all aspects of a project relevant to the largest numbers of stakeholders possible.

Providing sandboxes for limits-testing of new technologies at scale would present a particular challenge for any novel technology, given that such technology is always expensive and rare at the outset. The particularly technical aspects of quantum computing make it unlikely that it will become available in desktop devices any time soon. This means that access to quantum AI sandboxes will be more limited than for most technologies. Although quantum computing simulators that run on classical computers exist, and may be adequate for programmers to prototype algorithms, they would be wholly unsuitable for the kinds of limits testing we envisage since practical knowledge of the performance characteristics of quantum AI for many real-world applications will crucially depend on having systems which respond quickly and thus depend crucially on the parallelism of true quantum computing. Such real-time responsiveness cannot be simulated -- otherwise the quantum computer would be unnecessary to begin with! This likely discontinuity with previous technologies requires an even more careful application of practical wisdom at upstream stages (design, planning, and implementation) to avoid possible catastrophic failures of products once they enter the mainstream, a situation made only more pressing by the prospect of limited access to the machines for the majority of their development and deployment.

The provision of testing sandboxes with limited access is, nevertheless, preferable to having no capacity for testing at all. Obviously, for any kind of AI, whether based on quantum computing or not, testing requires the technical expertise of experts in the relevant technologies. Many obstacles to wise use of AI, from algorithmic bias to data privacy, are problems that depend on, even if they are not fully solved by, technical solutions delivered by experts (Kearns & Roth 2019). But limits testing of AI cannot be left solely to the experts. For one thing, the ethical and societal stakes of information technologies affect many more people than the experts at universities and technology companies, and those affected have a right to contribute to decisions. Additionally, it is often difficult for experts to predict how non-experts will use new technologies, or to predict how their behaviors will change (often in adverse ways) in response to new technologies (Bainbridge 1983). It is thus crucial to develop workspaces and sandboxes where these unexpected outcomes can be identified and mitigated before the real-world launch of an AI product.

This requires a degree of interplay between experts and non-experts which may be particularly difficult to achieve for quantum AI, since there are often additional barriers to understanding quantum mechanics without specialized training. This further discontinuity is introduced by the nature of quantum phenomena utilized in the function of quantum computers, including superposition and entanglement, that are necessary for making the manipulation of information more efficient (Murali et al. 2019). It is one thing (and surely no simple thing at that!) to ask a non-expert to understand the complicated workings of a simple artificial neural network, realized on a classical computer. It is quite another to ask them to understand the workings of a quantum computer, something even experts struggle with. It is thus not at all obvious how one should design and use quantum computers in a way that encourages wise use amongst the broadest number of users.

But why is this kind of deep knowledge necessary for wise use? One possible reply to our proposal is that competent and wise car drivers don't need to understand the workings of internal combustion engines, or their battery-driven alternatives. Operators should be able to get by using various heuristics that operate well in particular contexts without bringing the operator up to speed on the fundamental nature of quantum reality. We agree that, in general, utilizing such heuristics represents a rational, and often wise, way to manage one's limited cognitive resources (see Lieder & Griffiths 2020). But further knowledge becomes increasingly relevant at the limits. For instance, some knowledge of the difference between the two kinds of motor would be relevant to deciding which to rely upon in extremely dusty conditions. Likewise, the safest drivers are those who have driven cars beyond their limits in safe environments, and have thus good knowledge of the conditions under which (for example) tires will lose traction, and have the ability to recognize when they are in a situation not covered by the training provided by their prior limits-testing experience. The same principles apply to the design and use of quantum computers: although we cannot expect all users to have PhDs in physics, having a sufficiently tested and wisely designed product obviates the need for deep thinking at the margins.

Wise use of quantum AI involves exploration of use conditions which are hard to predict in advance, and even harder to predict without some level of basic knowledge about how the hardware and algorithms work. For instance, one important issue for some quantum algorithms is that reading out the result of the system interferes with it in ways that are not recoverable. If a readout happens too soon or too late, the result may be incomplete or inaccurate, but the probabilistic nature of quantum mechanics causes uncertainty about when to terminate a computation (Hagar 2003). Another example comes from the susceptibility of quantum computing to environmental noise, and the lack of robust error correction that is possible for quantum computing. This represents a sharp discontinuity with present technology, and presents a host of technical and ethical challenges for design: how can designers, companies, and regulatory entities be sure to manage the effects of a system that is unknowable in a way that current technology is not? We do not believe we know the answers to this question, nor do we think anyone could know them a priori. We do claim, however, that the framework of practical wisdom, and its focus on building breadth of knowledge and testing our knowledge at its limits, provides a uniquely suitable way of classifying, and attempting to solve, these problems.

Even if these problems can be handled reasonably under various conditions, the precise delineation of those conditions means that there are many unknown unknowns in quantum computing systems. This makes it hard to have a good grip on the limits of one's own understanding of those limits. In the context of AI, the differences between classical computing and quantum computing may be especially pernicious because of the tendency that many people have to anthropomorphically interpret the behavior and utterances of AI, supplying interpretations that attribute more intelligence than actually exists. More generally, the unintuitive role of non-classical probability theory in quantum mechanics may create new kinds of edge cases for quantum AI that do not exist for classical computing. As is the nature of edge cases and limits, we are not in an epistemic position currently to know (and be able to mitigate) all of the threats to wise and ethical use that might arise at the margins of quantum computing. If we adopt a practical wisdom framework at all levels of design and implementation, however, we will be in a position to identify threats from edge cases and minimize them as much as we can when they do arise.

In summary, though we remain noncommittal on the long-term future for quantum AI and skeptical about its short term significance, we believe that more analysis is both possible and desirable of the specific challenges that quantum AI presents to wise development. We do not pretend we know how this analysis will proceed; indeed, we think it is impossible to know the nature of many limits cases that might negatively impact individual lives and society at large before significant testing has been done. Our practical wisdom framework does, however, give one a way of approaching both the continuities and discontinuities in our ethical approaches to quantum computing. Some of those discontinuities rest on practical, but perhaps surmountable, limitations (e.g. the limited access many have to the hardware required to build quantum computers), while other discontinuities are inherent to the nature of the system itself (e.g. the inability to read the system without interfering). All the more reason, we think, to focus on

precisely delineating and understanding the limits of our knowledge, and building systems that take those limits into consideration when making important decisions about peoples' lives.

References

Bainbridge, L. (1983). Ironies of automation. In *Analysis, design and evaluation of man–machine systems* (pp. 129-135). Pergamon.

Hagar, A. (2003). A philosopher looks at quantum information theory. *Philosophy of Science*, 70(4), 752-775.

Karlan, B., & Allen, C. (in prep.) Engineered wisdom for learning machines.

Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43.

Murali, P., Baker, J. M., Javadi-Abhari, A., Chong, F. T., & Martonosi, M. (2019, April). Noise-adaptive compiler mappings for noisy intermediate-scale quantum computers. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems* (pp. 1015-1029).

Sevilla, J., & Moreno, P. (2019). Implications of Quantum Computing for Artificial Intelligence alignment research. *arXiv preprint* arXiv:1908.07613.

Theis, T. N., & Wong, H. S. P. (2017). The end of moore's law: A new beginning for information technology. *Computing in Science & Engineering,* 19(2), 41-50.