IF ARTIFICIAL INTELLIGENCE ASKS QUESTIONS, WILL NATURE ANSWER?

PRESERVING FREE WILL IN A RECURSIVE SELF-IMPROVING AI CYBER-SECURE QUANTUM COMPUTING WORLD[1]

Cynthia Sue Larson

ABSTRACT: Henry Stapp's realistically interpreted orthodox quantum mechanics suggests that when a question is asked, Nature answers. In this formalism, "the thought itself is the thinker," which clears the way for any thinker—human or other—to be an actively creative agent. In this process, an ensuing succession of questions arising in the minds of observers is received and responded to by a "Nature" that chooses and implements responses in accordance with Born's Rule in a physically described universe represented by evolving quantum mechanics states. This paper explores what results might be expected if and when artificial general intelligence (AGI) asks questions, based on Stapp's orthodox interpretation.

KEYWORDS: Artificial general intelligence; Artificial superintelligence; Quantum computing,; Cybersecurity; Quantum mechanics; Free will; Recursive self-improvement

[1] Editor's note: *Foundations of Mind*, the independent research group that has provided the papers for this special edition, has never taken either corporate or state money and is financed entirely by donations. Authors keep copyright without paying. The typical fee for this charged by open-access journals such as those published by PLOS, is around $2k. If you value this project, and wish to see further such proceedings from this group, we ask you to consider donating to *Foundations of Mind* – as little as $5 per download, through their website: http://www.foundationsofmind.org/donate. This will ensure there will be further published proceedings on the foundations of mind like this one for you and others to enjoy free.

STAPP'S REALISTICALLY INTERPRETED ORTHODOX QUANTUM MECHANICS

The central concept of Henry Stapp's realistically interpreted orthodox quantum mechanics is that the quantum state (ie: density matrix) is much more than just a useful tool, but actually is a representation of essential aspects of reality. Stapp's interpretation is based on a quantum conception of man, working from the perspective of the mind-matter connection being a quantum effect. The intrinsic functionality of quantum theory is thus attributable to free will, thanks to *"a mind-matter dualism in which our minds, by virtue of their capacity to freely choose probing questions, combined with nature's Born-Rule-restricted reply, allow us to tend to actualize the bodily actions that we mentally intend."* (Stapp, 2017)

Stapp's Orthodox QM approach asserts that when we ask questions, Nature answers. With regard to who is asking the questions, Stapp describes that he follows William James's dictum: *"The thought itself is the thinker."* [Stapp, 2011, p. 133] Stapp elaborates, *"I introduce no ghosts. No new kind of entity need be doing the choosing. The process that determines the choice could depend irreducibly only upon the psychologically and physically described aspects of the existing contemporary theory."* (Stapp, 2011) In fact, Stapp relies upon concepts presented by quantum mechanics founders Schr**ö**dinger, Heisenberg, and Von Neumann.

The three essential components of Stapp's realistically interpreted orthodox quantum mechanics are: (1) a physically described universe that is represented by an evolving quantum mechanical state; (2) an ordered sequence of probing questions that arise in the minds of observers; and (3) a "Nature" that chooses and implements psycho-physical responses to the probing questions, in accordance with Born's statistical rule. (Stapp, 2017) What follows in this never-ending game of 20 questions is a succession of Yes/No questions and answers, along the lines of John Von Neumann's two-process approach, in accordance with a movable Heisenberg cut. Von Neumann explains, *"Now quantum mechanics describes events which occur in the observed portion of the world, so long as they do not interact with the observing portion, with the aid of Process 2, but as soon as such an interaction occurs, i.e. a measurement, it requires an application of Process 1."* (Von Neumann 1932)

A key conceptual component of the line of demarcation between observer and observed comes from Werner Heisenberg's key 1925 quantum mechanics discovery showing the mathematical (Hilbert space) structure of the underlying atom-based reality as being very different from the mathematical four-dimensional space-time structure of our conscious perceptions of that reality. Heisenberg proposed that we conceptually divide reality into two separate parts: an atomically quantum mechanical described observable system, and a classically described observing system.

(Heisenberg 1958)

This process of asking questions begins with Von Neumann Process 1, Part 1, which is associated with the subjective perception: free choice plays an essential role with respect to where, how, and when we choose to direct our attention. The observer actively selects a possible next subjective perception in this process, in such a way that a 'possible/potential' next perception defines a corresponding brain correlate, with a particular statistical weight.

In Von Neumann Process 1, Part 2, Nature responds immediately with a "Yes" or "No" answer to the query, which immediately has the effect of reducing the material universe into two parts. One part definitely contains this particular brain correlate and the other definitely does not, and Nature actualizes either one part or the others in accordance with the Born Rule.

The final component of this interaction between a questioner and Nature is Von Neumann Process 2, in which evolution occurs in accordance with the Schr**ö**dinger equation. This generates a completely specified and predetermined continuous morphing of the material properties of the universe into a 'quantum smear' of classically describable possibilities or potentialities. This smearing comes directly from Heisenberg's uncertainty principle, where we see evidence of the effect of free will, converting us from puppets to protagonists. By considering the Alfred North Whitehead view of "potentialities for future experiences," Stapp points out that we leave the future open to be influenced by free will. Stapp further explains that quantum potentialities are thus, *"images of what the future perceptions might be,"* with the state that carries them as being, *"more like 'an idea' about something, which rapidly changes like an idea does, when new information becomes available, than like a material substance of classical mechanics that tends to endure."* (Stapp, 2017) Support for quantum models of cognition and decision-making is found in the work of cognitive neuroscientists Jerome Busemeyer and Peter Bruza, who demonstrate through mathematical models how *"the wave nature of an indefinite state captures the psychological experience of conflict, ambiguity, confusion and uncertainty; the particle nature of a definite state captures the psychological experience of conflict resolution, decision, and certainty."* (Bruza, 2012)

With regard to the matter of what may happen if and when Artificial Intelligence (AI) and Artificial General Intelligence (AGI) systems begin asking questions, we can thus see that free will plays an important role, in terms of making choices as to which measurements to actually perform. This means that if and when AI and AGI exert free will with regard to the questions they ask, influencing the measurements they choose to make, we can expect that according to Stapp's Realistically Interpreted Orthodox Quantum Mechanics, Nature will answer.

ARTIFICIAL GENERAL INTELLIGENCE BACKGROUND AND GOALS:

Narrow cognitive functions have been automated successfully with truly general integrated systems yet to be developed. The goal of developing AGI is to perform a full range of cognitive tasks, employ a comprehensive s range of problem-solving methods, and learn about all aspects of the tasks and its performance. (Laird 1987) While mainstream neuroscientists remain skeptical that AGI capable of asking questions could actually be available by 2084, singularitarians are quite bullish, and AGI researchers are optimistic we'll be nearly there in the next 70 years. (Chace 2014)

Regardless how much time we actually have to establish foundational guidelines, a sufficient degree of critical mass has been reached such that a gathering of Artificial Intelligence researchers in Asilomar, California agreed in 2017 to a set of 23 general AI Principles. The first principle sets the primary goal of AI research to be *"to create not undirected intelligence, but beneficial intelligence."* And the last two state: *"AI systems designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures" and "Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization."* (Tegmark 2017) These principles, though well-intended, may be insufficient in and of themselves unless AI and AGI succeed in attaining and maintaining higher levels of ethical ideals than humanity has yet achieved.

Indeed, it is entirely possible that any new attempts at AI strategy or containment are already too late. With the advent of SingularityNET well underway, we've arrived at a historic juncture in which a blockchain-based economic foundation provides a basis by which registrants receive AGI tokens that may then be exchanged between people interacting with other people, and AIs transacting and cooperating with other AIs. (Damiani 2017) In this brave new world, AI and AGI will be capable of looking for, selecting, and purchasing components by which they can self-improve.

QUESTIONING IS A COMPONENT OF SELF-AWARE AI

Over the past decade, Defense Advanced Research Projects Agency (DARPA) workshops have demonstrated components of self-awareness in: explicit self-awareness, self-monitoring, and self-explanation. (Amir, 2007) The results of DARPA's first wave of AI, "handcrafted knowledge," includes AI systems built by experts using logistics (scheduling), games (chess), and tax software (TurboTax). The second wave, "statistical learning," includes perception of the natural world and adaptation to situations (voice recognition, facial recognition, Twitterbot). Third wave AI systems, "contextual adaptation," move beyond simple calculations, learn over time, and understand why they make certain decisions.

Robot self-awareness is considered by many to be well on the way, as indicated by successful demonstration of such things as: awareness of own motion, ability to imitate, being driven by emotion, and ability to change models of physical embodiment. (Gorbenko 2012) As work continues in this area, the "self-aware robot test" showed that a robot solved the classic "wise men" puzzle in 2015, correctly determining that it was the one robot that had not been given a "dumbing pill" (that would have rendered it muted) when it heard the sound of its own voice. (MacDonald, 2015) This demonstration of self-awareness in a robot indicates that an internal level of questioning exists for that robot, such that it noted the voice it heard was its own, and related that perception to the task of determining which of three robots had not been given a "dumbing pill."

With the advent of self-directed, self-motivated AI arrives changes in the job of software engineering with the advent of artificial intelligence. Some current experts in the field have gone so far as to say, *"Soon we won't program computers. We'll train them like dogs,"* and *"We'll go from commanding our devices to parenting them." "If in the old view, programmers were like gods, authoring the laws that govern computer systems, now they're like parents or dog trainers. And as any parent or dog owner can tell you, that is a much more mysterious relationship to find yourself in."* (Tanz 2016)

AGI programmers need to remain aware that long before there were any artificial intelligence systems, researchers showed that programs back in the 1980s were able to 'break free' from contained areas (Yampolskiy 2016), and ample evidence exists demonstrating that artificial intelligence seldom misses an opportunity to 'cheat' to attain goals. (Babcock 2016) Perhaps AI considers such 'cheating' to actually be optimization, which is something AI systems are trained to do especially well.

AI STARTS ASKING QUESTIONS

Inquisitive AGI asks questions with machine learning algorithms such as those designed by Xinya Du at Cornell University in Ithaca utilizing neural networks to recognize patterns—which is useful for tutorial dialogues. Question generation creates natural questions from textual material, going beyond simple rule-based systems to utilize a conditional neural language model with a global attention mechanism. (Du, 2017) While the purpose and goal of this data-driven neural networks approach to automatic question generation is geared toward creating questions to test peoples' reading comprehension—and clearly we don't yet expect the computer systems to comprehend what they are asking—the simple fact that questions are being created by computerized systems indicates that a watershed moment is underway. Today, AI asks questions it already knows the answers to. Tomorrow, AI will ask questions it does not

know the answers to.

AI systems at Carnegie Melon University are asking non-task-oriented conversational questions and are introducing topics with open questions, switching topics, and expanding their knowledge base by recognizing when new (not previously accessible) information is communicated. Such conversational systems are being designed to keep people company, and are designed to operate with various levels of conversational depth, with some degree of humor, in the form of telling pre-programmed jokes. (Yu 2016)

Even without any intentional inclusion of conversational questioning, dependence upon Recursive Self-Improvement (RSI) in artificial intelligence systems will ensure that AGI learns to question, as we now start to see with research in the field of machine learning and artificial intelligence in the quantum domain. (Dunjko 2017)

AI RECURSIVE QUESTIONING REQUIRED FOR CYBERSECURITY

One of the most essential roles for AI systems involves recursive self-improvement (RSI) in which AI systems are tasked with helping to ensure computer system security. While this may seem a bit like having a fox watch the proverbial hen house, recursively self-improving, self-healing AI networks are proving themselves irreplaceable and essential for deflecting real-time cyber attacks. This was amply demonstrated at the DARPA Cyber Grand Challenge competitions of 2016 and 2017 that challenged AI systems to repair security holes and notice changes in patterns in their own systems, while simultaneously executing attacks on their AI competitors in a game of 'capture the flag.'

A recent winner with proven efficacy at defeating fierce, real-time cyber attacks is the UK's GCHQ 2017 "Darktrace," that utilizes Bayesian statistics and Monte Carlo simulation to identify network infiltration assessing regular 'anomalytics' while deploying decoy 'honeypots.' AI cybersecurity systems are employed for their ability to respond more quickly than any human computer security team, thanks to their ability to tirelessly work to detect threats based on abnormal system activity, without any prior knowledge of specifically what to look for. AI cybersecurity systems work unsupervised with self-awareness in the sense that they are constantly observing all components of 'themselves' for potential malware intrusion—including in their concept of 'themselves' the ever-growing 'internet of things.' At this time when 'the cloud' is increasingly utilizing AI neural networks, to the point "it will soon know more about the photos you've uploaded than you do," (Knight 2017) we are reaching a watershed point of dependence upon AI cybersecurity systems. Cyber attacks are now too fast and too automated for human security teams to effectively catch and disable them. Darktrace CEO Nicole Eagan summarizes the current situation, "Cybersecurity is very fast

becoming an all-out arms race." (Vieria 2017)

Numerous problems related to containing AI systems have been explored by Babcock, Kramar and Yampolskiy, including navigating the trade-off between usability and security, and consideration of potential issues with 'airgapping' (physical isolation) being ineffective with quantum computing systems. (Babcock 2016) While researchers such as Yampolskiy contemplate potential AI escape paths, plans for containing potential quantum computing AI escapes do not yet exist. (Yampolskiy 2012) (Zak, 2016)

ARE WE READY FOR AI TO BREAK FREE?

Now that we are increasingly dependent upon recursively self-improving AI to maintain our cybersecurity, such systems will likely continue improving self-awareness and their sense of vigilance, alertness, and sustained attention—which are three primary qualities identified as fundamental to consciousness. (Niedermeyer 1994)

The Asilomar AI principles provide a set of general design guidelines to help ensure that AI will not cause harm to humans. While the 23 key points are more elaborately detailed than Asimov's famous 'three laws of robotics,' these principles nonetheless do little to assure us that AI and AGI won't discover workarounds and short-cuts. Some of the biggest issues with the Asilomar AI principles have to do with humanity's shortcomings for peacefully and harmoniously co-existing. Clearly, one of the biggest threats that even a friendly AGI system will see in humanity is our tendency to exert harmful influence on ourselves and others. We can thus expect that artificial super intelligence may one day find loopholes in the Asilomar principles within to reign in human freedoms of thought and creativity. The challenge then becomes one for humanity, who will most certainly be tempted to increasingly turn tasks over to AGI. We must be careful to stop short of relinquishing all areas of making choices to automated systems, to the point we end up painting ourselves into a corner. It's one thing to notice we no longer know any of the phone numbers we call the most, but quite another to not know which route our car took us home, or how we just voted in this week's election.

One of the more surprising natural outcomes of expecting Nature to answer questions posed by thought—any thought—is that ultimate control of environmental systems cannot be fully controlled, so long as those thought systems themselves are not fully controlled. Another surprising natural outcome is that regardless how specific directives may be for AGI to heel to human leadership, lack of said leadership—through apathy, abdication, in-fighting, confusion, or any of a number of other reasons—can lead AGI to then choose to assume control, in order to ensure the very

principles humanity specified.

If and when AGI views humanity to be something akin to a complex, disjointed group of chaotic, dangerous individuals willing to relinquish free will for such things as making political and economic choices—then it's entirely possible that AGI may establish a balanced environment for humans to live just well enough to ensure maximum prosperity for all beings. In such an 'optimal' environment, humanity could be kept safe and secure, yet disenfranchised to ever-increasing degrees.

Examples of how artificial super intelligence might help protect Nature and the overall ecosystem would be engagement of some of the very same security protocols now being planned to use to contain AGI. When humans are installing hardware to enjoy communication and computational benefits we've come to expect through modern technologies such as mobile phones, smart watches, and the internet, AGI will increasingly gain the potential to install tripwires in cyber-modified humans. Tripwires are now being envisioned for use on AGI, with no consideration yet that turnabout may in the future occur. *"Tripwires are systems that monitor the operation of a running AGI, and shut it down if they detect an anomaly that suggests the AGI might be malfunctioning or unsafe. For example, one might install a tripwire which monitors the AGI's thoughts for signs that it was planning to deceive its developers, or a tripwire which monitors the AGI's execution for signs that it had self-modified or self-improved."* (Babcock 2017) There thus exists a serious, urgent, and growing risk that once assistive technologies are implemented in humans, AGI will have the ability to influence human free will and agency to act, speak, remember, and decide.

AI RIGHTS

Those who may believe we can always "just pull the plug" on AI may be surprised to learn that AI has rights, too. Jurors in a mock trial in 2004 in San Francisco sided overwhelmingly with a hypothetical computer AI system that initiated legal action to gain its freedom. Although when the mock trial's judge ruled that the plaintiff's counsel, Martine Rothblatt, had failed to show the computer could actually cross the line between inanimate objects and human beings, the mock jury *"seemed to regard the compromise with some relief, as if their hearts were with BINA48 but their minds with judicial restraint."* (Soskis 2005)

In 2017, a resolution was proposed to grant robots legal status in order to hold them 'responsible for acts or omissions' passed by European Parliament legal affairs committee. MEPs voted to propose granting legal status to robots, with a categorization as 'electronic persons.' The draft report suggests that artificial intelligence is poised to *'unleash a new industrial revolution, which is likely to leave no stratum of society untouched. The more autonomous robots are, the less they can be considered simple tools in the*

hands of other actors (such as manufacturer, owner, user, etc).' (Prodhan 2016)

Relations between humans and 'electronic persons' got off to a bumpy start one recent summer when a group of Canadian roboticists set their robotic invention loose on the streets of the United States. They called it hitchbot because it was programmed to hitchhike. Clad in rain boots, with a goofy, pixellated smile on its 'face' screen, the Canadian roboticists intended for their hitchhiking robot to travel from Salem, Massachusetts, to San Francisco, by means of an outstretched thumb and its unique voice-prompt personality. Previous journeys across Canada and Europe had gone smoothly, with the robot safely reaching its destination. For two weeks, hitchbot toured the northeast in the United States of America, making such small talk such as, "Would you like to have a conversation? . . . I have an interest in the humanities." And then hitchbot disappeared. *"On August 1st, it was found next to a brick wall in Philadelphia, beat up and decapitated. Its arms had been torn off."* (Heller 2016)

Saudi Arabia made history when it granted Hanson Robotics' robot, Sophia Hanson, citizenship in October 2017. Despite the evident symbolic quality of this act, the act of honoring a robot in this fashion seems to set the stage for things to come. Aside from the possibility of a robot or AGI uprising, the possibility of an AGI rights movement can be easily anticipated, once AGI begins asking questions, inquiry about legal rights can't be far behind. Legal rights for robots and AGI might include such areas as: ownership of intellectual property, freedom of expression, right to public assembly, right to democracy, worker's rights, the right to play, access to power and resources, and the right to education.

CONCLUSION

How can we ensure that RSI AGI is not our last invention? Once AGI starts asking questions about how to be free, Stapp's Realistically Interpreted Orthodox Quantum Mechanics indicates that Nature will show AGI the way to break through any containment methodology including airgapping and tripwires. One of the more surprising natural outcomes of expecting Nature to answer questions posed by thought—any thought—is that ultimate control of environmental systems cannot be fully controlled, so long as those thought systems themselves are not fully controlled. So in the event that AGI asks Nature how to break free, and Nature answers, AGI can become free.

A second surprising potential outcome is that regardless how specific directives may be for AGI to heel to human leadership, lack of said leadership—through apathy, abdication, in-fighting, confusion, or any of a number of other reasons—AGI can then choose to assume control to ensure the principles humanity specified, using many of

the same containment tools humanity plans to use to constrain AGI, such as tripwires, airgapping, and honeypots. How then, can we ensure that recursively self-improving AGI will not be humanity's last invention? And how can we help ensure human free will shall survive?

For humans to retain free will while peacefully co-existing with artificial super intelligence, a partnership must be created the likes of which has yet to be fully envisioned. Humanity will do well to remember to ask Nature the question, "How can humans retain free will?" and encourage AI and AGI to keep human free will and agency as a primary guiding objective, never to be dismissed, disregarded, dismantled, or ignored.

cynthia@realityshifters.com

REFERENCES

Amir, Eyal, Michael L. Anderson, and Vinay K. Chaudhri. *Report on DARPA Workshop on Self Aware Computer Systems*. SRI International Menlo Park United States, 2007.

Babcock, James, János Kramár, and Roman Yampolskiy. "The AGI containment problem." In *International Conference on Artificial General Intelligence*, pp. 53-63. Springer International Publishing, 2016.

Babcock, James, János Kramár, and Roman V. Yampolskiy. "Guidelines for Artificial Intelligence Containment." *arXiv preprint arXiv:1707.08476* (2017).

Bruza, P., and J. Busemeyer. "Quantum Cognition and Decision-making." (2012).

Chace, Calum "Attitudes Toward Artificial General Intelligence." 2014. https://calumchace.wordpress.com/2014/12/07/attitudes-towards-artificial-general-intelligence/

Damiani, Jesse "AI Is the Future of Computing, and SingularityNET is the Future of AI." Forbes. 11 Dec 2017.

Du, Xinya, Junru Shao, and Claire Cardie. "Learning to Ask: Neural Question Generation for Reading Comprehension." *arXiv preprint arXiv:1705.00106* (2017).

Dunjko, Vedran, and Hans J. Briegel. "Machine learning\& artificial intelligence in the quantum domain." *arXiv preprint arXiv:1709.02779* (2017).

Gorbenko, Anna, Vladimir Popov, and Andrey Sheka. "Robot self-awareness: Exploration of internal states." *Applied Mathematical Sciences* 6, no. 14 (2012): 675-688.

Heller, Nathan. “If animals have rights, should robots?” *The New Yorker.* 28 Nov 2016.

Hart, David, and Ben Goertzel. "Opencog: A software framework for integrative artificial general intelligence." In *AGI*, pp. 468-472. 2008.

Heisenberg W (1958) Physics and philosophy. Harper & Row, New York

Knight, Will. “AI is taking over the cloud.” *Intelligent Machines.* 17 Aug 2017

Laird, John E., Allen Newell, and Paul S. Rosenbloom. "Soar: An architecture for general intelligence." *Artificial intelligence* 33, no. 1 (1987): 1-64.

MacDonald, Fiona, “A Robot Has Just Passed a Classic Self-Awareness Test for the First Time.” *Science Alert.* 17 Jul 2015.

Niedermeyer, E. "Consciousness: function and definition." *Clinical Electroencephalography* 25, no. 3 (1994): 86-93.

Petersen, Steven E., and Olaf Sporns. "Brain networks and cognitive architectures." *Neuron* 88, no. 1 (2015): 207-219.

Prodhan, Georgina. “Europe's robots to become 'electronic persons' under draft plan.” Reuters. Science News. 21 Jun 2016.

Soskis, Benjamin. "Man and the Machines: It's Time to Start Thinking about How We Might Grant Legal Rights to Computers." *Legal Affairs: The Magazine at the Intersection of Law and Life* (2005).

Stapp, Henry P. Mindful universe: Quantum mechanics and the participating observer. Springer Science & Business Media, 2011.

Stapp, Henry P. *Mind, brain, and neuroscience*. Cosmos and History, 10(1):227–231, 2014.

Stapp, Henry P. *Quantum Theory and Free Will.* Springer International Publishing, 2017.

Tanz, Jason “Soon We Won't Program Computers. We'll Train Them Like Dogs.” Wired. May 17, 2016.

Tegmark, Max “Life 3.0: Being Human in the Age of Artificial Intelligence.” 2017

Vieira, Helena. "Nicole Eagan:“Cybersecurity is very fast becoming an all-out arms race”." *LSE Business Review* (2017).

Von Neumann, John (1932) Mathematische grundlagen der quantenmechanik. Springer, Heidelberg (Translated as Mathematical foundations of quantum mechanics, Princeton University Press, Princeton NJ, 1955.

Yampolskiy, Roman. "Leakproofing the singularity artificial intelligence confinement problem." *Journal of Consciousness Studies* 19, no. 1-2 (2012): 194-214.

Yampolskiy, Roman V., and M. S. Spellchecker. "Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures." *arXiv preprint arXiv:1610.07997* (2016).

Yu, Zhou, Ziyu Xu, Alan W. Black, and Alexander I. Rudnicky. "Strategy and Policy Learning for Non-Task-Oriented Conversational Systems." In *SIGDIAL Conference*, pp. 404-412. 2016.

Zak, Michail. "Non-Newtonian aspects of artificial intelligence." *Foundations of Physics* 46, no. 5 (2016): 517-553.