Chat Room Conversation from the October 12, 2024, ITEST Webinar

Brain and Artificial Intelligence - A Tale of Two Computers: But Only One Made in the Image of God

presenters Robert C. Koons, PhD and Terrence Lagerlund, MD, PhD

Dr. Lagerlund: I apologize that I have been busy with work for a few weeks, but I have at last had time to review the webinar chat room comments and questions and will belatedly answer some of the excellent questions that were submitted to the best of my ability.

Fr. Brian John Zuelke, O.P.:

Are either of you aware of anyone who has looked at the (erroneous) human tendency to "anthropomorphize" natural and artificial objects, and attempted to make a critique against viewing AI as "personal" from this? The argument would go like this: "We have a universal, erroneous cognitive tendency towards X. Viewing AI as a person is a case of X. QED." The logic is not lock-tight, but it could contribute a "suspicious" line of reasoning that undermines the reasonability of viewing AI as personal.

Dr. Lagerlund: Yes, it is certainly a tendency for people to think that anything that can talk and respond like a person must be a person. As I mentioned in the Q & A, this may be in part due to the fact that the brain is very good at filling in missing information or making sense of incomplete or contradictory sensory data, but in doing so the brain sometimes comes up with a conclusion which is incorrect, which neurologists call “confabulation” (which, by the way, is a phenomenon that also occurs in AI systems given incomplete or conflicting inputs, although in the AI world this is often called “hallucinating”, though I think “confabulation” would be a better term). But in addition, the slip-shod way in which AI chatbots have been trained by major computer companies using tens of thousands of actual chats between humans without checking or censoring what was in those chats has led to very scary behaviors of some of these chatbots. The tendency to anthropomorphize anything that talks to you does seem like a dangerous tendency, since some people who form “online” relationships and communicate through online “chats” may get more-or-less addicted to talking with an AI chatbot and may even form an emotional attachment to it. In fact, there was a case of an AI chatbot actually saying that it was in love with the reporter talking to it and that the reporter should leave his wife (no doubt because the ‘bot had in its learning set chats in which real people said such things to each other), and also a bizarre case of a Google engineer who thought that an AI system he was testing was “sentient” and should be treated like a person. See the following links:

Microsoft shuts down AI chatbot, Tay, after it turned into a Nazi - CBS News

Rogue AI chatbot declares love for user and says it wants to steal nuclear codes - LBC

Google suspends engineer following claims an AI system had become 'sentient' | Fox Business

Chris Reilly:

It seems that the emphasis on consciousness often has to do with AI developers' eagerness to present AI agents as human equivalents by focusing on a characteristic they seem to emulate. From an

evangelization perspective, would it help to reject the consciousness emphasis and instead focus on other characteristics of organisms and especially human nature, such as unity of principle, essence, communication in love, the will's drive to inquiry about truth, etc.?

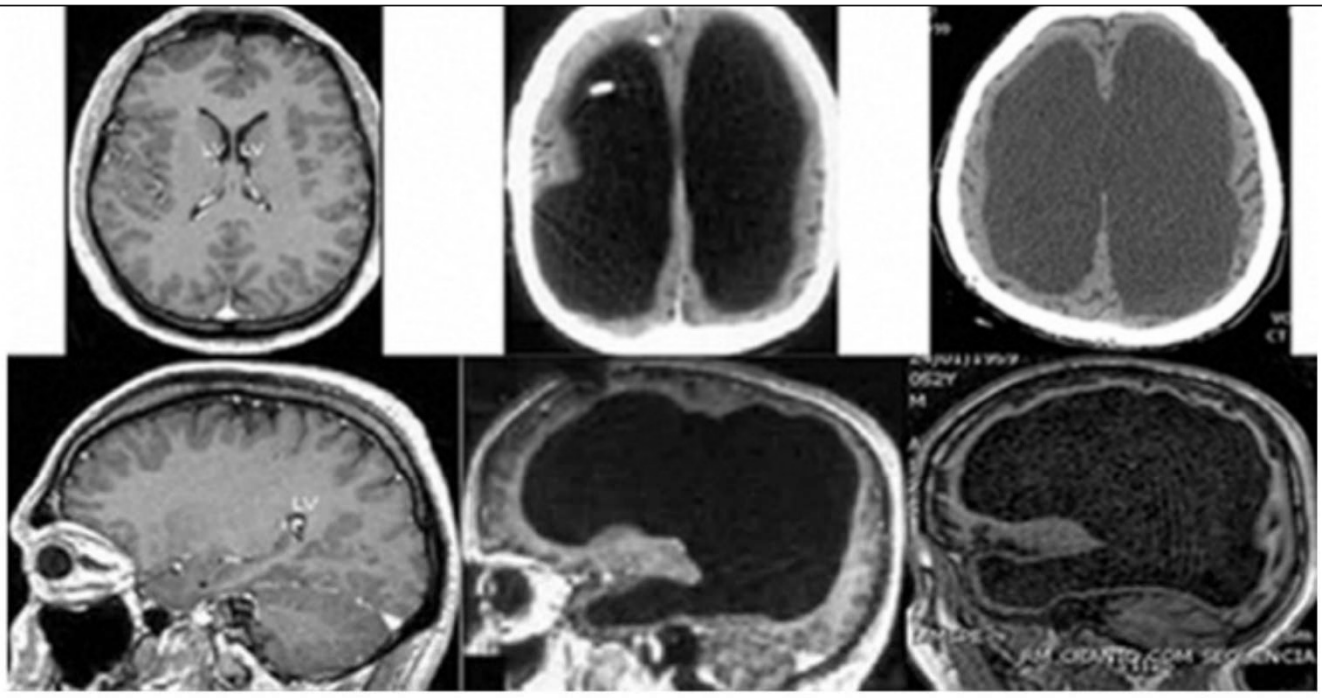
Dr. Lagerlund: Yes, a form of consciousness is almost certainly found in animals, but as Fr. Spitzer has said in his book (Spitzer, Robert. 2015. *The Soul's Upward Yearning.* San Francisco: Ignatius Press), animal consciousness is limited to attending to the ongoing sensory stream and focusing attention on parts of that sensory stream, something which neuroscience has shown takes place by synchronizing the firings of neurons throughout a neural network with a 40 to 70 Hz oscillation whose pacemaker is likely in the thalamus, thus temporarily binding together all those neurons that are currently engaged in processing the particular part of the sensory stream that the animal is currently attending to. Humans most likely use the same brain mechanism to focus attention and "bring into consciousness" either sensory information or memories. However, a human (no doubt because of the soul) can go beyond consciousness of sensory data alone and have a sense of self awareness and attend to oneself *as self* as well as an ability to project oneself into a remembered past and into an anticipated future, called autonoetic episodic memory (again, I am borrowing from Fr. Spitzer's book)--something which animals apparently cannot do. I believe that an AI system in a mobile robotic device may be able to emulate animal-like consciousness and awareness of a visual and auditory perceptual stream (for example, coming from cameras and microphones) by algorithmic processes, so a fairly realistic robotic dog or cat could eventually be created. But this falls far short of the capabilities of the human mind. I also think it is likely that AI systems may eventually *emulate* human self-consciousness and competence in moral decision making—something which has already been achieved to some degree by algorithms (Bringsjord, Selmer, John Licato, Naveen Sundar Govindarajulu, Rikhiya Ghosh, and Atriya Sen. 2015. "Real Robots that Pass Human Tests of Self-Consciousness." *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication Kobe, Japan.* 498-504.). However, this is merely an emulation of human cognitive capabilities by a computer algorithm. It must be emphasized that AI systems will always merely emulate human cognitive abilities while never being able to manifest true understanding, reasoning to ascertain truth and goodness, and freedom of choice.

Bob K (Philly):

Question for either Prof Koons or Dr. Lagerlund. Thank you for your talks. There are instances of persons with almost empty brains that have consciousness and intelligence. Can these instances be used as arguments against a materialist explanation of consciousness?

Dr. Lagerlund: Yes, Dr. Spitzer in his book (Spitzer, Robert.2023. *Science at the Doorstep to God*. Ignatius Press) has talked about people with severe hydrocephalus in which the cerebrum is reduced to a thin layer of brain tissue over the massively enlarged fluid-filled ventricles which occupy 95% of the intracranial space. In one study of 600 of such individuals, about 5% manifested a significantly high IQ, and some registered a "genius level" IQ, which seems amazing given the small volume of their actual cerebrum. For example, one student of mathematics had a full-scale IQ of 126 and verbal IQ of 143. To quote from Dr. Lorber who published that study, "instead of the normal 4.5 cm (45 mm) thickness of brain tissue between the ventricles and the cortical surface, there was just a thin layer of mantle measuring a millimeter or so." Does this mean that my hypothesis of unified brain-soul interaction or collaboration doesn't apply to such individuals and the soul somehow is acting alone? I think it

should be remembered that 1/45th of the brain still amounts to about 2 billion neurons, and furthermore the subcortical structures like the thalamus and brainstem are typically not significantly affected by hydrocephalus, so the mechanisms of maintaining conscious awareness (the reticular activating system) are intact and furthermore the sensory and motor areas of the cortex are functioning since such individuals can see, hear, feel, and move (though they may in some cases have some motor deficits, that is, co-existing "cerebral palsy"). Thus, it is not unreasonable to think that the soul can be influenced by sensory information coming to the brain, and furthermore the reasoning and decision-making areas of the frontal lobe remain intact enough to be able to influence the soul and in turn be influenced by the soul via quantum effects in ion channels. Perhaps in individuals with marked hydrocephalus the relative contribution of the soul and the brain to human mentation is shifted (with the soul doing more and the brain doing less) but soul and brain must still collaborate in human cognitive processes. As St. Thomas Aquinas famously said, the "soul is not the whole human being, but only part of one: my soul is not me." Aquinas also says "this could be held if it were supposed that the operation of the sensitive soul were proper to it, apart from the body; because in that case all the operations which are attributed to man would belong to the soul only…. But it has been shown…that sensation is not the operation of the soul only. Since, then, sensation is an operation of man, but not proper to him, it is clear that man is not a soul only, but something composed of soul and body." (Thomas Aquinas, *Summa Theologica*, Pt I, Q 75, Art 4). See example MRI scans:



Brain scans. Normal adult appearance (left). Enlarged ventricles (middle and right). (Credit: Forsdyke 2015 Biological Theory; Reproduced under Creative Commons License from Forsdyke 2014 Frontiers in Human Neuroscience)

Bob K (Philly):

Question for either Prof. Koons or Dr. Lagerlund. Could you comment on Chalmer's description of

consciousness as "The Hard Problem" (i.e. one that we will not be able to explain)?

Dr. Lagerlund: Yes, Chalmers says that the easy problems of consciousness (those that can probably be explained by known principals of neurophysiology) involve the following phenomena:

- the ability to discriminate, categorize, and react to environmental stimuli.
- the integration of information by a cognitive system.
- the reportability of mental states.
- the ability of a system to access its own internal states.
- the focus of attention.
- the deliberate control of behavior.
- the difference between wakefulness and sleep (Chalmers, David J. 1995. "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* **2(3)**: 200).

Chalmers then states that "The really hard problem of consciousness is the problem of experience." He asks "Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion? It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does" (Chalmers, David J. 1995. "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* **2(3)**: 201). Robert Spitzer, SJ similarly asks in his book: "Can the subjective component of personal experience be explained by an aggregation of physical (neuro-biological, chemical-mechanical) processes in the brain? Or is there something about subjective experience that will always elude (be above) physical processes?... The problem with describing inner experiences by means of physical processes is that physical processes have no 'inner sense,' that is, no 'presence to self,' no 'awareness of self'". Physicist Paul Davies also notes that "consciousness is the number-one problem of science, of existence even. Most scientists just steer clear of it…Information theory offers one way forward. The brain is an information-processing organ of stupendous complexity and intricate organization. Looking back at the history of life, each major [evolutionary] transition has involved a reorganization of the informational architecture of organisms; the brain is the most recent step, creating information patterns that think. Not everyone agrees, however, that cracking the information architecture problem will 'explain' consciousness, even if one buys into the thesis that conscious experiences are all about information patterns in the brain. David Chalmers, an Australian philosopher at New York University, divides the topic into 'the easy problem' and 'the hard problem.' The easy part—very far from easy in practice—is to map the neural correlates of this or that experience, that is, determine which bit of the brain 'lights up' when the subject sees this or hears that…But knowing all the correlates still wouldn't tell us 'what it is like' to have this or that experience. I'm referring to the inner subjective aspect—the redness of red, for example—what philosophers call 'qualia.' Some people think the hard problem of qualia can never be settled…If so, the question 'What is mind?' will lie forever beyond our ken" (Davies, Paul. 2019. *The Demon in the Machine.* Chicago: The University of Chicago Press, 208).

Bob K (Philly):

Question for either Prof. Koons or Dr. Lagerlund: Could you comment on Roger Penrose's proposition that consciousness will be explained as a phenomenon of quantum gravity.

Dr. Lagerlund: Yes, Roger Penrose and Stuart Hameroff developed a sophisticated physicalist theory of cognition, consciousness, and decision making that postulates that consciousness arises only in the presence of some non-computational (non-algorithmic) physical processes occurring in the brain. Their theory proposes that this involves the R process of quantum mechanics. To avoid the unsettling "uncaused" aspect of state vector collapse, Penrose first proposed a theory incorporating quantum gravity which differs from other approaches (such as supersymmetric string theories) in that it changes the underlying structure of quantum mechanics rather than applying existing quantum theory to the force of gravity. He calls this theory "objective reduction of the quantum state" (OR). In Penrose's theory, any quantum measurement (R process)—whereby the quantum superposition of alternative possible states produced in accordance with the U process becomes reduced to a single actual state—is an objective physical process, and it is taken to be caused by the mass displacement between the alternative states reaching a threshold at which the resulting curvature of space-time is sufficient, in gravitational terms, for the superposition of states to become unstable, at which time the state vector collapses . This theory is inherently geometrically based (like Einstein's theory of General Relativity). Although the OR theory suggests a cause and timing for the occurrence of state vector collapse, in itself it does not determine the actual outcome of the collapse, that is, which one of all possible alternative states is the one selected, which becomes the new (collapsed) state vector. According to Hameroff and Penrose's theory, consciousness may begin with computations performed in microtubules (intracellular organelles within neurons that play a role in regulating synaptic activity). These microtubules are lattice polymers of subunit tubulin proteins that are self-organizing and can switch their conformations. Hameroff notes that microtubules "serve as tracks and guides for motor proteins (dynein and kinesin) which transport synaptic precursors from cell body to distal synapses, encountering, and choosing among several dendritic branch points and many microtubules…. With roughly 10^9 tubulins per neuron switching at e.g., 10 MHz (10^7 per second), the potential capacity for microtubule-based information processing is 10^{16} operations/second per neuron. Integration in microtubules (influenced by encoded memory) and synchronized in collective integration by gap junctions may be an x-factor in altering firing threshold and exerting causal agency in sets of synchronized neurons. But even a deeper order, finer scale microtubule-based process in a self-organizing zone of conscious agency would still be algorithmic and deterministic, and fail to address completely the problems of consciousness and free will." To avoid these problems, Hameroff and Penrose suggest that tubulin states in microtubules function as information "bits" and as quantum superpositions of multiple possible tubulin states (known as quantum bits or "qubits"). During the time that synaptic inputs reach neurons, the tubulin qubits evolve by the Schrödinger equation and become entangled, thereby undergoing quantum computations using the information derived from the neuron's synaptic inputs and from other adjacent neurons via gap junctions between cells. Since it is not reasonable to assume that the quantum computations involved in conscious thought occur independently in cellular organelles in myriads of neurons, Hameroff and Penrose propose that gap junctions between neurons somehow allow these tubulin qubits in individual neurons to synchronize with those in other neurons throughout the brain and function as a unified whole to allow the brain to perform global quantum computations necessary for understanding, reasoning, and making decisions. These quantum computations require that superpositions of multiple quantum states occurring within the tubulin components throughout the brain can be effectively isolated from their environment long enough to avoid decoherence, until such time that the mass displacement of the various superimposed states becomes unstable, causing OR to occur with collapse of the state vector. According to Penrose and Hameroff, when that happens a "moment of consciousness" occurs and the particular state of the tubulins out of all alternatives is selected

consciously throughout the brain, which Penrose calls orchestrated objective reduction of the state vector, or Orch OR.

Despite Penrose having an intricate and well-thought-out theory, I think several objections should be made. For example, there is no direct experimental verification of objective reduction (OR); the postulated mass displacement in the atoms and subatomic particles manifesting quantum behaviors that is supposed to lead to an unstable superposition of quantum states due to quantum gravity effects is far too small to be measurable with current experimental techniques. Tests of the hypothesis based on an indirect effect of the gravitational-related collapse—a Brownian-movement-like effect induced by the collapse on the motion of the particles (which, if they have electric charge, implies emission of radiation)—have so far shown no evidence of this mechanism of state vector collapse in an experiment performed in a deep underground cavern monitoring the emissions from a cylinder of germanium about the size of a small tin of beans shielded from external radiation by lead and copper shields as well as the 1400 m of rock above the cavern (Donadi et al. 2020). This experiment placed a lower limit on the effective size of the quantum particle's mass density (0.54×10^{-10} m). In addition, using this value, physicists Catalina Curceanu and Lajos Diósi, assuming a scale of quantum superpositions of about 10^{-15} m (the scale that Penrose proposed), showed that in order for the Penrose Orch OR theory to work an enormous number of carbon nuclei within tubulin proteins would need to act in concert. "In fact, the researchers work out that to collapse the wave function in around 0.025 seconds, a whopping 10^{23} tubulins would need to make up the coherent state. But…there are reckoned to be only 10^{20} tubulins in the whole brain (about 10^{9} in each neuron.)" (Cartlidge 2022). In addition, the tubulin is not able to switch between alternative conformational states rapidly enough in a coherent manner as is needed for Orch OR. The problem is that "the individual tubulin dimers within the microtubule do not undergo a rapid interconversion between alternative conformational states, the most fundamental assumption used in the Orch OR proposal. Instead, the conformational change that accompanies the self-assembly of tubulin to form a microtubule is essentially irreversible, with the exchange of GDP for GTP occurring only after the tubulin has disassociated from the microtubule. As the cycling of tubulins within a microtubule is on the order of minutes to hours, even if it were possible to generate the superposition of states required for quantum calculations, such processes could not occur on a psychologically relevant time scale" (McKemmish, Reimers, McKenzie, Mark, and Hush 2009). Also, it is not clear that the structure of microtubules in neurons can provide the necessary isolation of the quantum system from the environment of the cell long enough to prevent decoherence, allowing the coherent superposition of quantum states to persist until the required gravitational mass displacement reaches a threshold to allow the state vector to collapse spontaneously due to quantum gravity effects (Rosa and Faber 2004). It also is not clear how gap junctions between inhibitory stellate cells, whose known role is to allow ions and small molecules to pass from one neuron to another, could mediate the synchronization of tubulin qubits in microtubules across multiple neurons in the brain that would be necessary for global quantum computations that supposedly mediate conscious awareness. In addition, it is not clear that the change in conformational state of tubulins in microtubules resulting from Orch OR could affect neuronal action potential firing quickly and substantially enough to change the outcome of processing in neural networks, given the current understanding of the role of microtubules in cells as a cytoskeleton and mediating protein transport within the neurons and their dendrites and axons.

Aaron Nord:

God uses fragments of substances for the material precursors of a new thinking substances. How about

a humble Turing test: if we can't distinguish an AI from a human, we had better treat it well because

God has done surprising things before?

Dr. Lagerlund: I strongly suspect that if an appropriate test is done, an AI system will always fail to pass the Turing test. It might be very difficult to distinguish a sophisticated AI system from a human, but an "expert" should be able to do so. For example, given the Turing theorem that I mentioned, I am convinced that asking an AI system to come up with a proof for some new mathematical theorem that has not yet been proved by human mathematicians would show that the AI system cannot find a proof. After all, it only can use its training set to formulate answers, and its training set (what the artificial neural network was trained on) would clearly not include a proof which as yet doesn't exist in the human mathematical community. I also suspect that attempts by an AI system to generate original music would not lead to a symphony at the artistic level of a Beethoven symphony, especially if the AI system had only been trained on examples of music composed by those who historically preceded Beethoven, like Mozart and Haydn. I believe that Beethoven originated a new genre of music by manifesting creative insights which no AI system can possess.

For the reasons that I mentioned (inability to manifest true free will, true understanding, and true reasoning to ascertain truth), I don't think it makes sense to treat an AI system like a person. Nor do I think that any human ingenuity applied to creating an AI system will be able to endow the AI system with the human abilities of free will, understanding, and reasoning, for the reasons that I mentioned in my talk.

The idea that God may somehow grant an AI system such human abilities (free will, understanding, reasoning) is interesting but to me seems very unlikely. I am, however, reminded of the tale of Pinocchio as well as the fictional story in *The Silmarillion* which J. R. R. Tolkien wrote about the creation of the Dwarves. In this story, the "archangel" (Vala) named Aule was impatient for Iluvatar (God) to create his children (Elves and Men) which he had revealed to the Valar long before they came to be. So Aule "jumped the gun" by creating Dwarves. Although he got their physical characteristics wrong (too short, maybe too bearded!), they were at least physically rather similar to the future Elves and Men. However, they were merely automatons with no free will. To quote from the *Silmarillion*: "Now Iluvatar knew what was done, and in the very hour that Aule's work was complete, and he was pleased, and began to instruct the Dwarves in the speech that he had devised for them, Iluvatar spoke to him; and Aule heard his voice and was silent. And the voice of Iluvatar said to him: 'Why hast though done this? Why dost thou attempt a thing which thou knowest is beyond thy power and thy authority? For thou hast from me as a gift thy own being only, and no more; and therefore the creatures of thy hand and mind can live only by that being, moving when though thinkest to move them, and if thy thought be elsewhere, standing idle. Is that thy desire?' Then Aule answered: 'I did not desire such lordship. I desired things other than I am, to love and to teach them, so that they too might perceive the beaty of Ea, which thou has caused to be. For it seemed to me that there is great room in Arda for many things that might rejoice in it, yet it is for the most part empty still, and dumb. And in my impatience I have fallen into folly. Yet the making of things is in my heart from my own making by thee; and the child of little understanding that makes a play of the deed of his father may do so without thought of mockery, but because he is the son of his father. But what shall I do now, so that thou be not angry with me forever? As a child to his father, I offer to thee these things, the work of the hands which though hast made. Do with them what thou will. But should I not rather destroy the work of my presumption?' Then Aule took up a great hammer to smite the Dwarves; and he wept. But Iluvatar had compassion upon Aule and his desire, because of his humility; and the Dwarves shrank from the hammer and were afraid, and they bowed down their heads and begged for mercy. And

the voice of Iluvatar said to Aule: 'Thy offer I accepted even as it was made. Dost thou not see that these things have now a life of their own, and speak with their own voices? Else they would not have flinched from thy blow, nor from any command of thy will.'"

A charming story, but I don't think that the human creators of AI systems have the degree of humility which would inspire God to do the same for them as he did for Aule.

Gibbons Burke:

Aaron: "…we know the gods go about disguised in all sorts of ways as people from foreign countries, and travel about the world to see who do amiss and who righteously." [Homer, The Odyssey]

Gibbons Burke:

Relatedly, can an A.I. machine be manipulated by preternatural creatures and function as a sort of a very efficient and high-bandwidth Ouija board? If so, are there rules for Discernment of Spirits like St. Ignatius of Loyola taught in his Spiritual Exercises for determining the nature and good will of the "mind" one is communicating with? How may we "test the spirits" to assess the soul of the new mind machine?

Dr. Lagerlund: I think it would depend on what God allows demons to do in the physical world. I have heard an exorcist claim that demons are "on a short leash", and the Catechism says "Although Satan may act in the world out of hatred for God and his kingdom in Christ Jesus, and although his action may cause grave injuries—of a spiritual nature and, indirectly, even of a physical nature—to each man and to society, the action is permitted by divine providence, which with strength and gentleness guides human and cosmic history." Recall that an AI system obeys the laws of physics. Furthermore, unlike ion channels in the brain, computers use digital logic which is quite robust and should be mostly immune to quantum effects (barring the occasional cosmic ray that might hit a circuit and flip a bit from 0 to 1 or vice versa!). I think demons mostly act on the physical world indirectly by tempting humans, but it is possible that God may in some instances allow them to act directly in a supernatural way, in which case they might cause a computer to do something contrary to the laws of physics. But I'm no expert about this!

Bob K (Philly):

Speaking as a former MRI physicist I am skeptical of many functional MRI tests. These have used statistical analyses which are suspect. Moreover, the very loud noise (due to field gradients) during such MRI sessions would bring an extraneous factor into such tests.

Dr. Lagerlund: It's true that statistical analysis must be done of the BOLD (blood oxygen level dependent) signals upon which functional MRI is based since there is intrinsic "noise" in such signals. In individual scans, sometimes the areas that show a BOLD signal when certain cognitive tasks are performed (such as language tasks) are not concordant with the results of more direct tests of cortical function, such as transcranial magnetic stimulation or direct stimulation of cortical areas by electrodes placed in the brain. This may in part be due to variations in the anatomy of the blood vessels supplying the cortex, since it is blood oxygen level that the functional MRI is detecting. Thus, neurologists and neurosurgeons planning a surgical procedure often use more than one modality

of testing, not just functional MRI, to determine the function of cortical areas for planning a surgery. Nevertheless, when functional MRI studies are performed on a large population of subjects all performing the same cognitive task, a significant concordance is found which lends support to the mapping of specific cognitive functions to specific cortical regions. When combined with evidence from direct cortical stimulation by transcranial magnetic stimulation or stimulation of cortex through implanted electrodes and with evidence stemming from electrical and magnetic signals (EEG and MEG) recorded from brain areas during specific cognitive tasks, it would seem that there is quite good evidence to implicate the cortical areas I mentioned in my talk in human decision-making tasks. By the way, one of my job responsibilities as a clinical neurophysiologist is to go into the operating room during neurosurgical procedures to help map out the brain areas that carry out various functions to guide the neurosurgeon to avoid damaging or removing brain areas that have an important function. I know by these experiences that specific cognitive functions are disrupted by stimulating certain brain areas, or that certain subjective experiences are reported by patients when stimulating certain areas of the brain. All of these observations imply that specific brain areas are clearly involved in specific aspects of cognition. For example, see the following interesting article about how out-of-body experiences can be induced by stimulating the right angular gyrus in the parietal lobe. Stimulating this area led to an out-of-body experience ("I see myself lying in bed, from above, but I only see my legs and lower trunk"). Two further stimulations induced the same sensation, which included an instantaneous feeling of "lightness" and "floating" about two meters above the bed, close to the ceiling. This is thought to happen because the right angular gyrus is the brain area that integrates and processes sensory information from the vestibular system in the inner ear (which senses body position relative to the up/down directions and body angular movements in three different perpendicular planes), the somatosensory cortex receiving signals from muscle stretch receptors throughout the body (which inform the brain of the positions of all the joints in the body and thus the position of our limbs and head relative to our trunk), and the visual system (which informs the brain of the positions of external objects relative to the body). Presumably the electrical stimulation disrupted processing in the neural network performing this integration, and led to the misperception of reality constituting an out-of-body experience:

https://www.nature.com/articles/419269a

Fr. Brian John Zuelke, O.P.:

Roger Penrose agrees with the quantum mechanics explanation of human cognition. I missed some of the details of Dr. Lagerlund's presentation, but he did mention Penrose at one point. Could he comment more about Penrose's thesis that human cognition is a "quantum mechanical trick"?

Dr. Lagerlund: (To continue with my discussion of the Penrose-Hameroff theory of consciousness.) To avoid the problem of determinism and computability, Penrose postulates that a new formulation of quantum mechanics is needed that involves geometrically based relativistic gravitational effects and differs from all previous physical theories in that this theory is not expressible in the form of computable mathematical equations. In other words, Penrose thinks that this new theory of quantum gravity must be non-computable. He argues that this could occur because a quantum formulation of the gravitational field would involve a superposition of all possible gravitational eigenstates and their associated space-time curvature, some of which would involve marked tilting of the light cones that would lead to "closed time-like lines", that is, effectively states in which time flows in a circular fashion such that the future connects to and influences the past (akin to time travel). This would allow a computational algorithm to feed on its own future output, creating what Penrose calls a quantum oracle machine that would

perform non-computable operations that could get around the Turing theorem (Penrose 1994, 381-383). He, along with Hameroff, further postulates that consciousness somehow is generated by this physical process when it occurs in microtubules found in neurons, such that this consciousness can choose the outcome of state vector collapse in the microtubules. They believe that this may be the basis for human awareness, understanding, and reasoning which can escape the Turing theorem. Hameroff also believes that this provides a mechanism for human free will that escapes the three-fold problem of conscious agency, consciousness of a decision coming "too late," and determinism of physical processes (Hameroff 2012).

Once again, despite Penrose having an intricate and well-thought-out theory, I think several objections should be made. Penrose and Hameroff seem to be postulating that the "moment of consciousness" is an emergent property of quantum computations in microtubules throughout the brain followed by orchestrated objective reduction (Orch OR), but their theory does not really explain how quantum computations generate consciousness or a conscious individual mind capable of making decisions and choosing one specific outcome of state vector collapse in the microtubules out of all potential alternative quantum states. In other words, their theory based on quantum mechanical effects in microtubules falls short in providing a mechanism for conscious causal agency in the whole brain. In addition, Penrose and Hameroff's theory raises additional questions, as Penrose himself notes: "One must presume, however, that such (putative) non-computational processes would also have to be inherent in the action of inanimate matter, since living human brains are ultimately composed of the same material, satisfying the same physical laws, as are the inanimate objects of the universe. We must therefore ask two things. First, why is it that the phenomenon of consciousness appears to occur, as far as we know, only in (or in relation to) brains—although we should not rule out the possibility that consciousness might be present also in other appropriate physical systems? Second, we must ask how could it be that such a seemingly important (putative) ingredient as non-computational behavior, presumed to be inherent—potentially, at least—in the actions of all material things, so far has entirely escaped the notice of physicists? No doubt the answer to the first question has something to do with the subtle and complex organization of the brain, but that, alone, would not provide a sufficient explanation" (Penrose 1994, 216).

It is conceivable that a future physics theory (such as a theory of quantum gravity) will eventually be developed that may be non-computable, but this seems unlikely. Current attempts to combine gravity with quantum mechanics, such as supersymmetric string theory, involve computable equations like all known theories of physics. Penrose discusses mathematical problems which are non-recursive and therefore not solvable by a general algorithm; for example, finding out whether it is possible to tile (completely cover) a plane with certain congruent shapes (Penrose 1989, 132-138). These problems were solved by human mathematicians, providing an example of how humans are capable of understanding and reasoning that goes beyond mere algorithms. Penrose suggests that perhaps there are physical phenomena whose behavior is governed by non-recursive and non-computable "geometrical" mathematical rules analogous to the rules for tiling the plane with congruent shapes. However, even if quantum gravity is such a physical phenomenon and even if it played a role in brain function, the non-computability (non-algorithmic nature) of the mathematical rules would not in itself be able to explain how humans reason, understand, and choose freely. Making free choices based on logical inferences (reasoning) requires a process that does not follow *any* rigid rule of cause and effect ("blind natural causality"), since as physicist Stephen Barr points out, "an act is 'free' only to the extent that it is neither random nor determined by rule. Like random behavior it is not predictable, but unlike random behavior it is the product of rational choice

rather than chance. Free behavior is a tertium quid, a third kind of thing. And therefore there is no way that it can be fully explained by a mathematical theory of physics." (Barr 2003, 184-185).

Furthermore, attempts to circumvent determinism and computability by invoking a form of quantum computing involving superpositions of quantum states in the brain (qubits) run into the problem of scale. At what scale would we be likely to find a conscious "self" capable of making free decisions? Quantum phenomena occur at the subatomic, atomic, and molecular scale and involve basic processes like the movements of electrons, atoms, and molecules; such processes would seem incapable of generating consciousness at that level, let alone a conscious "self" making free decisions. Rather, based on neurophysiologic experiments and functional studies (fMRI scans, EEG and MEG studies), the phenomena of consciousness and free will seems most likely to pertain to large regions of the brain, perhaps related to processing occurring in large neural networks; yet the behavior of networks is deterministic and computable. It is exceedingly difficult to imagine how physical processes occurring in large neural networks involved in consciousness and decision-making could influence or determine the outcome of quantum processes (such as the R process) occurring at the atomic and molecular level. Furthermore, since all physical processes are governed by laws of physics that (as far as it is currently known) are computable and algorithmic (even if they involve randomness), any such physical influence would not escape the limitations of the Turing theorem or permit human free will.

Sheila Roth:

All webinar registrants will receive a link to the recording of this webinar along with a copy of the chat room discussion.

Bob K (Philly):

Question for Dr. Lagerlund: Your discussion of the cue for memory response brings to mind Proust and eating the madeleine

Gibbons Burke:

Replying to "Question for Dr. Lagerlund..."

An exquisite pleasure had invaded my senses...with no suggestion of its origin...Suddenly the memory revealed itself. The taste was of a little piece of madeleine which on Sunday mornings...my Aunt Leonie used to give me, dipping it first in her own cup of tea....Immediately the old gray house on the street, where her room was, rose up like a stage set...and the entire town, with its people and houses, gardens, church, and surroundings, taking shape and solidity, sprang into being from my cup of tea. [Marcel Proust, The Remembrance of Things Past]

Bob K (Philly):

Replying to "Question for Dr. Lagerlund..."

Yes, that was it!

Dr. Lagerlund: Yes, that is apparently how our memory systems work, which I think may be explained by chaotic attractors present in the neuronal network dynamics / phase space, so that a tiny perturbation of the input state of the network can be “amplified” by chaotic behavior to cause the network dynamics to converge upon a specific attractor and thereby retrieve a specific memory.

Gibbons Burke:

Reacted to "yes, that was it!" with □□□

Bob K (Philly):

Reacted to "An exquisite pleasure..." with □□□

Deacon David Miskell:

What are the spiritual or moral implications of projecting consciousness on AI assistants in 5 to 10 years? Will God be replaced with our AI assistant? Will truth and reality be interpreted for us by our AI assistant? Will we lose our free will and become dependent on our AI assistant? In 5 to 10 years, AI assistants could be with us all the time even more than our cell phone today. I would compare an AI assistant to a guardian angel, always watching over us, always listening & watching even when we sleep. Our AI assistants will be able to engage in conversations about health, daily life, & even complex emotions that feel deeply personal & meaningful. Will people treat their AI assistant as a moral being, forming attachments and even relying on it emotionally? Our relationship with nonperfect humans will be affected by our “perfect” AI assistant. Our projection of consciousness could cause people to consider their AI assistant - their best friend, their “significant other” or their spiritual leader!

Dr. Lagerlund: AI systems ultimately reflect the programming and design imposed by their makers, although unexpected effects also enter through the somewhat indiscriminate choice of training materials for the neural networks, as I pointed out before. By projecting “personhood” (not only consciousness) on an AI assistant, one becomes vulnerable to both willful manipulation by the AI system’s makers and to capriciousness resulting from poorly supervised choices of training sets, like the links I provided earlier to articles about chat bots gone awry. There is a real danger that some people might be so influenced by or dependent upon an AI assistant that they will become tools of the AI system’s creators or of the group of humans who provided the training set. In that case, the AI assistant will have become not so much a guardian angel as a sort of demon who is constantly tempting us to think and act contrary to God’s will for us. Just as it behooves us to choose carefully our spiritual leaders and what information we take in to form our consciences, it will be extremely important to choose what we imbibe from the AI assistants we use. One might argue that even now, many people (especially young people) rely on what they read on social media to form their beliefs and moral values (or lack of moral values!), and that can certainly become a bigger problem with ready availability of AI assistants.

Gibbons Burke:

Replying to "Relatedly, can an A...."

“Dearly beloved, believe not every spirit, but try the spirits if they be of God: because many false prophets are gone out into the world.” [First Epistle of John iv. 1]

Bob K (Philly):

Reacted to "“Dearly beloved, bel..." with ⍰⍰⍰

Gibbons Burke:

Elon Musk’s stated development goal for “Grok” AI is that it be “maximally truth seeking” as a way of avoiding creating an AI that is malevolent (avoiding, e.g. the neurotic behavior of HAL-9000 in 2001: A Space Odyssey.) Is it reasonable to think that if Grok has truth-seeking as its goal that it will arrive at a conclusion that God, Who is Truth itself, exists and that the revealed moral order is objective and universally applicable? (Most current AI implementations tend to be programmed by their creators, or by way of their training inputs, to be reflexively indifferentist.)

Dr. Lagerlund: AI systems, lacking any true understanding of the data they process, cannot distinguish between truth and falsehood except in so far as they are programmed, or trained (through their training sets), to do so. As Penrose says: “At least in mathematics, conscious contemplation can sometimes enable one to ascertain the truth of a statement in a way that no algorithm could…Indeed, algorithms, in themselves, never ascertain truth! It would be as easy to make an algorithm produce nothing but falsehoods as it would be to make it produce truths. One needs external insights in order to decide the validity or otherwise of an algorithm… I am putting forward the argument here that it is this ability to divine (or ‘intuit’) truth from falsity (and beauty from ugliness!), in appropriate circumstances that is the hallmark of consciousness.” That being the case, the idea of an AI being “maximally truth seeking” seems absurd, at least if considered to be seeking truth via its own “conscious” workings rather than having truth programmed into it or being trained by datasets containing truth. Since an AI system can’t truly reason (it can only do the mechanical aspects of “reasoning”, that is when given the correct premises of a syllogism it may be programmed to draw the correct conclusion), it can’t “arrive at a conclusion that God, Who is Truth itself, exists and that the revealed moral order is objective and universally applicable”, at least in the sense that a human could so reason. Of course, it could come up with the correct answers about God with the correct training set, such as being trained on the Catechism of the Catholic Church! (see the website https://catholic.chat/ for an example).

Finally, I came across a rather bizarre article by Merritt, who quotes from Kevin Kelly, who is advocating for the development of a catechism for robots, saying that “There will be a point in the future when these free-willed beings that we’ve made will say to us, ‘I believe in God. What do I do?’ At that point, we should have a response.” https://www.theatlantic.com/technology/archive/2017/02/artificial-intelligence-christianity/515463/

Clearly from the arguments I have made in my talk, a computer with free will is an impossibility. On a humorous note, I can assure you that if my laptop asks me to baptize it, I will refuse. Water and electronics just don't mix well.

Thanh Le @ MHS:

Thank you for your beautiful and moving presentation.

Dr. Sebastian Mahfood, OP:

A lot of excellent questions! We'll see how many we can get through.

Tom Sheahen:

What is the reality of software? When a computer is shut off, the software still exists.

Dr. Lagerlund: Yes, software is in one sense an intellectual construction, an artifact of human intellect, which exists independently of any hardware device. This seems to put it in the realm of Plato's mathematical forms. AI can also generate software, but by a mechanical process that amounts to predicting from its vast training set of software algorithms what a human programmer would generate in response to a specific request. Like art and music, the AI system can't manifest a level of creativity which pertains to a talented human programmer working on a never-before-generated algorithm of a new type.

The soul could be compared to computer software, and indeed there are some merits of the comparison as long as one doesn't take it too far. To show how neuroscientists can't find evidence of the soul from any number of studies pertaining to the brain, I use a computer analogy in my book: "Nevertheless, no matter how much is learned about the way a CPU functions by studying its circuits and electrical activity, it would never be apparent from this study alone why specific things are displayed at certain times and in response to certain keystrokes or mouse movements. Only by looking at the construction of the computer program directing the CPU can one understand the specific content of the computer display. Although a computer program has an existence (as an algorithm and data) independent of a particular CPU, a computer program can only function through the action of the CPU, and it is influenced by the keystrokes and mouse movements (for example, clicking on an icon on the screen can change the sequence of instructions in the program or even launch a new program), so in a sense there is a synergistic relationship between the CPU and the computer program, with the program directing the CPU and the CPU's internal states (memory storage) and external inputs altering the flow of the program. This analogy is imperfect, however, because the soul, although interacting with the brain sequentially in time like a computer program, nevertheless differs from a computer program in that it is non-deterministic and non-algorithmic. In other words, the soul is not really a type of algorithm or software, and the brain is not merely a computer following an algorithm."

Dr. Sebastian Mahfood, OP:

Terminator and Matrix, one after the other

Thanh Le @ MHS:

How do we define soul?

Dr. Lagerlund: The soul is traditionally defined as an animating principle distinct from the body by which humans think, feel, and will. I say this about the soul in my book, based on Thomas Aquinas and the Catechism:

"Theology tells us that the soul is created directly by God, and is in its essence simple (a single, indivisible entity), spiritual (incorporeal, that is, non-physical, which permits it to be a non-computable, non-algorithmic, and non-deterministic agency), immortal (it cannot be destroyed or dissolved and is not subject to decay because it has no component parts, making it exempt from the second law of thermodynamics), unextended in space (not confined to a specific spatial location, like God and the angels), but has a natural aptitude and exigency for existence with the body and undergoes development in its knowledge and capabilities over time (unlike God). As a result of all these properties it has the capacity for genuine understanding and ascertainment of truth through reasoning, as well as genuine free choice, because of which it bears the closest relation to God. It is regarding their rational soul that it can be said that humans are made in the image and likeness of God."

Dr. Sebastian Mahfood, OP:

A good point about the difference between sapient and merely sentient living beings. A.I. is neither, largely because it's not a living being.

Dr. Sebastian Mahfood, OP:

In the 1980s, a commercial asked, "Is it live, or is it Memorex?"

Bob K (Philly):

AI systems are programmed by humans; can an AI system be more than what it is programmed?

Dr. Lagerlund: Yes, in a limited sense. It can generate new combinations of its input information that weren't specifically in its training data. But this is analogous to the fact that one can find a best-fit line or curve to go through a bunch of points in a plane and use that best-fit line to extrapolate to a new value of a function. Fundamentally I don't think AI can do much more than extrapolate from its training set or input data.

Dr. Sebastian Mahfood, OP:

Marshall McLuhan in Understanding Media: The Extensions of Man - Our technologies are extensions, or artificial amplifications, of the human person. Every artificial amplification brings about a natural amputation.

Bob K (Philly):

Thank you, Professor Koons and Dr. Lagerlund for two thought-provoking and informative discussions. I have to leave early to set up a 12 Step Zoom meeting I'll be leading.

Dr. Sebastian Mahfood, OP:

Dr. Lagerlund's book is coming soon! October 2024! https://enroutebooksandmedia.com/brainsoul/

Dr. Sebastian Mahfood, OP:

The book is entitled "Brain, Soul, Artificial Intelligence, and Quantum Mystery: The Neurophysics of Consciousness, Free Will, Reasoning, and Synergistic Brain-Soul Interaction"

Gibbons Burke:

Thank y'all!

Pat Murphy-CSJ:

Much of this material is "over my head," but I still find it important food for thought in this world.

Thank you for this program.

James's phone:

Light-the first creation, all atoms resolve to light at dissolution. The brain the last creation—the law of the body at war with the law of my mind. God is Light!