An Experiment: The Gods of Cyberanthropology

In this paper I discuss some ways of considering the 'programming' of humans through neuroanthropology as a pathway into understanding the programming of Artificial Intelligence. A.I. could 'evolve' into a new type of life that may well achieve some level of self-awareness. I discuss the problems of ethics and religion as they apply to human and A.I. behavior and end with four fundamental questions that any self-aware programming will have to have built in or discover on its own. These are the questions of: Meaning (including purpose), Sin (value hierarchy), Pain (or pleasure), & Death.

The issues regarding A.I. as a potentially self-aware entity have not been comprehensively addressed or have been addressed in such a way as to imply there are still good choices open to us as humans. We often assume that A.I. isn't intractable or 'wicked' – but it does present itself as a 'wicked' problem (Churchman 1967). As we seek to understand more about Artificial Intelligence (A.I.), many different streams of thought can cast light onto the nexus between Man and Program. One of these streams of thought is neuroanthropology, the integration of neuroscience and anthropology, first coined by graduate students Juan Dominguez Duque and Paul Mason and adopted by Daniel Lende and Greg Downey to start what has become a whole new scientific field of endeavor. (Lende, D. 2012)

Neuro-anthropology might just look like a fancy way of saying 'nature' and 'nurture'; the 'nurture' component being Anthropology and the 'nature', Neuroscience. In this article, this nature-nurture dichotomy is just as interesting for programming A.I. as it is for 'programming' humans. In this paper, human 'programming' will fall under neuroanthropology; programming Artificial Intelligence (A.I.) will fall under its cousin: Cyberanthropology, i.e. the nature and nurture of A.I. This paper will lead to a number of relevant questions that we need to answer regarding cyberanthropology in order to understand it better. We will then go on to explore ethics and morality in A.I. and look at how little we really understand about how it might develop.

Setting the Ground Rules

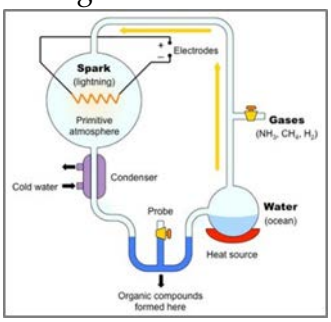
Too often, even scientists are 'only human', which might raise doubts about their suitability for the complex philosophical and introspective issues this paper attempts to address. Be that as it may, this humanness has many benefits, but also brings biases and pitfalls that should not be overlooked.

In my estimation, one core neuroanthropology bias is more glaring than others and it should serve as a sign of humility in the face of great unknowns. This bias manifests itself in the assumptions about ethics and origins of life and we should knock it about first before diving more deeply into the subject matter.
For example, it should be a humbling fact that much of the literature in neuroanthropology simply assumes classical Darwinism got it all right (Dobzhansky, March 1973) and that the brain seems to modify itself based on a Niche Theory (Lende & Downey, 2012) that suggests that morality and ethics come from cultural tradition, information transfer, and development of social skills. But in adopting this viewpoint, neuroanthropologists have made a tremendous 'god of the gaps' error.

"God of the Gaps" means: a spurious explanation of anything not nailed down by science; usually reserved for heaping hot coals on Evangelical Creationist preachers. The neuroanthropology mistake is to apply Evolutionary Darwinism to everything except for the *actual* beginning of life: the encoded DNA molecule. In other words, DNA and life are just assumed and the assumption is grounded in Classic Darwinism. This error manifests itself once we carefully look at the nature of DNA; more specifically, it becomes more glaring when we consider the nature of *information*. We will now touch on this subject briefly because it will have implications for the entire paper.

BANG – and there you have it…

Most scientists accept the 'fact' that Miller and Ury put the difficulties of 'origin of life' to bed with their amino acid experiments in 1953 (Miller & Ury, 1953). Of course, they didn't, but their superficial evidence is rather widely held as conclusive and given popular credence by authors such as Dan Brown in his latest book "Origin" (Brown, 2018).



In Brown's book, the billionaire explorer of life-origins takes a new look at the actual Miller and Ury test tube in 2003 – 50 years after the original experiments. In the original test tubes, he finds *new* combinations of amino acids and puts the difference between the number of old 1953 amino acids and the new 2003 amino acids in an Artificial Intelligence program. Using the 50 years extrapolation, the program goes through billions of simulated years and predictably the A.I. calculations come up with what we know now as DNA. As the main character allows the program to continue past the supposed onset of DNA toward an unknown theoretical future, it finds that a new lifeform is on the evolutionary horizon! Brown calls this new lifeform: Technium. Brown's 'Origin' is fiction of course, but the linear, logical approach to the origins of life he paints is simple and seductive and colors much of science today.

Nevertheless, there is evidence that Darwin's step-by-step approach for the explanation of life would have needed much more time than the Universe has ever had access to (Nagel, 2012). Fred Hoyle, in his article on the origins of life (Hoyle, 1981) suggest that the chances of life (read: DNA) spontaneously popping up were about 10^{4000} to 1. Since the number of atoms in the observable universe is approximately 10^{60} to 10^{85}, Hoyle's assertions are simply a scientist's way of saying: 'impossible'. To use the paleontologist Stephen J. Gould's well-worn aphorism, DNA's appearance "is a '*just so*' story"; a tip of the hat to Rudyard Kipling's "Just So Stories" for children of 1902, which include 'How the Leopard got his spots'. It is 'just so' that a provable impossibility happened. And because this provable impossibility happened, Darwinian Evolution was able to pick up the pieces and retain its step-by-step credibility. Consequently, the Darwinian starting point assumed by much of the relevant science seems to be fundamentally flawed and Gould would assert that *all* of Evolutionary Psychology is too weak a framework to hold the story of neuroanthropology, let alone cyberanthropology (Gottlieb, 2012).

Another blow to Darwinian 'step-by-step'-ism can be found in Stuart Kauffman's highly acclaimed book, "At Home in the Universe" (Kauffman, 1995) where he presents, not only the sheer technical impossibility of Darwin's basic theory, but counters with clear evidence of amazingly catalyzed 'explosions' of new life variants. In Kauffman's words: "We need to paint a new picture" of how earthly species came to be (Kauffman, 1995, p. 9). He goes on to talk of 'vast veins of spontaneous order' such as the structure of snowflakes. He shows that their magnificent 'six-fold order' is the way it is because it is 'that kind of thing'. In his words, we do not know where every atom is, but we know the properties of a snowflake because it is 'that kind of thing'. Lende and Downey lean, too, in this direction: "…powerful but overly-simple models of evolution… need to give way [in order to] provide richer accounts that incorporate data emerging from genetics, paleoanthropology, comparative neuroscience, and anthropology…". (Lende, 2012, p. 125)

Not to worry, I will not attempt to take the path of Creationism or Intelligent Design here, but when the foundations of a scientific theory like Darwinism are questioned and shown to be shaky, one might want to proceed with extra caution. This is what we will attempt to do; proceed cautiously knowing the bridge spanning this chasm has a number of rotted planks. Kauffman again: "We appear to have been profoundly wrong. Order, vast and generative, arises naturally." This is an intriguing insight, but still begs the question of why a universe like ours would have just that sort of property, that it

is just that finely tuned, that there is 'something' rather than 'nothing'; but let's all just agree that the appearance of DNA and Life is *just that sort of thing*. In this way, we can move on 'as if' classical Darwinism has a role to play in helping us unravel the complexities of nature and nurture in humans and in cybernetics.

So, having agreed that the 'thunder' of Life is attributable to Thor, we can move on to employ Evolution as 'that sort' of mechanism that might be the hand of God and might be a fundamental property and might be both. Regardless, we will use the concept in an "as-if" way to better understand how humans are building the brain of A.I.; nurturing A.I. – potentially the next new life-form.

Human Programming

Having settled our 'as if' assumptions, there is much being done to understand the evolution of the human brain on our way to the A.I. brain. For this article, I will use Lende and Downey's work ("Evolution of the Brain", 2012) as my main source of insight since it is most purely from the new discipline of neuroanthropology (Lende, 2012). As Clare Kelly and Hugh Garavan have asserted: "Research with both animal models and humans has shown that changes in neural representations can be induced not only in response to lesions of input or output pathways, but that the organization of the adult cerebral cortex can change substantially as a result of practice and experience" this can result in increases (strengthening), redistribution (adapting), and decreases (efficiency) in neural functions (Kelly & Garavan, 2004).

There seem to be three main areas or theories of influence on the human brain that determine its size and ability. I've chosen these as the most potentially fruitful areas of discourse for theorizing about cyberanthropology.

First, there is a Social Intelligence Hypothesis which "focuses on the demands and opportunities that being a long-lived, highly social species bring and the role of competition with our own kind". In other words, how our brains' size and capabilities is shaped by the challenges we face and the life stages we go through. The function of neuroplasticity or the brain's (in-) ability to conform to its challenges is a key element.

Second, there is the Brain Development Niche model. This "incorporates ideas about cultural tradition, information transfer, and the development of technological, foraging, and social skills". This seems closely related to the work of Eric Trist in the Sociotechnical School of Human Relations (Trist, 1951). Perhaps the most famous example from Trist is that of miners at the coal face having the work of drilling, blasting, filling, and carting of coal replaced by an enormous, noisy, but efficient machine. The traditions of how teams worked together, how they used and transferred information, how their careers – not to mention their days – developed, and how they adapted to the new social environment would have profound effects on their nervous systems.

1. **Social Intelligence Hypothesis**: development from the demands and opportunities of the environment
2. **Brain Development Niche model**: development arises from tradition, information transfer, as well as technological, foraging, and social skills
3. **Sensory, Motivational, & Body Changes Model**: developmental changes affect body functions and brain functions

Finally, the Sensory, Motivational, and Body Changes Model shows the evolutionary and developmental function of the brain and demonstrates that not only the brain is changed, but the body, as well. A prime example of this is Erin P. Finley's work "War and Dislocation: A Neuroanthropological Model of Trauma among American Veterans with Combat PTSD" (Finley, E. 2012). Finley suggest there are numerous triggers that physically modify the brain and body such as stress, horror, dislocation, and grief. Reaction times and accomplishment of physical tasks is quicker in these stressed individuals, but learning is negatively affected including performance on memory tests. The body responds by physically re-wiring the brain as is also the case in cocaine addicted rats (Siviy, et. al.

2014) or 'attention density' in change-stressed employees (Rock, 2006); the threat or stressful situation may resolve itself and the physical addiction may dissipate, but the triggered and enhanced neurons have now moved from a flicker of desire to a roaring cascaded of activated need.

The overlaps between these three models are vast, implying a Grand Unified Theory might be around the evolutionary corner. Using these three areas of Neuroanthropology, I will now look at how we can understand what has and will happen in Cyberanthropology as A.I. 'grows up' and its 'Brain' develops.

Cyberanthropology

The promise of Artificial Intelligence (A.I.) or Artificial General Intelligence puts most other technologies in a prehistoric light. However, "These technologies (like Microsoft 'Tay') may perpetuate cultural stereotypes" (Johnston, 2016) or throw up other undesirable effects. At Microsoft, the online A.I. exposed to all and anything internet users had to offer, learned that weapons and insects were "unpleasant", but flowers and musical instruments were "pleasant", even, presumably, accordions. This seemed to be a 'good' outcome, but like a pet cockatoo, online A.I.s are going to be at the mercy of the teenagers who use their goofy ideas and partial insights to populate them.

There are many instances of innocuous A.I. missteps that just need a bit of tweaking such as translating 'gender free' words in one language into 'gendered' words in another. But, in the MIT Technology Review, William Knight looks at the next step in A.I. which relies on 'deep learning' (Knight, 2017). These new algorithms are making decisions based on variations of calculations on immense blocks of data that defy human capabilities to understand. Not only is the amount of information humanly indigestible, it is beyond human comprehension just how the A.I. program makes some of its fundamental decisions based on the information. In his words, "There's already an argument that being able to interrogate an AI system about how it reached its conclusions is a fundamental legal right". This may seem to be the logical next step after teaching the 'parrot' of A.I. to swear at your neighbor's wife.

Like the cascade of human neurons that fire to create what we experience as consciousness, the A.I.s using deep learning also employ levelled-learning. This is a process where the first set (level) of simulated neurons gets a highly pixelated or rough sketch of something. Subsequent deeper levels learn to recognize profounder layers of detail and abstraction. "It might be part of the nature of intelligence that only part of it is exposed to rational explanation. Some of it is just instinctual."

For those who do not easily recognize existential dangers: that last quote points out that A.I. might have 'instincts'. Let that color how you read the rest of this document.

At Google, the Deep Dream A.I. began to come up with images of 'reality' that looked distinctly like it had taken LSD. Its recognition algorithm produced bizarre animals and shapes that, in some ways, were reassuringly recognizable and in other ways terrifying. Things, like a dumbbell, were represented with a human arm attached (an arm that wasn't actually there) because the A.I. had perceived the two things as one and categorized the arm and the dumbbell as one. These mirages are eerily akin to how humans fill in missing information as well as how humans dream.

Whether the A.I. is supplying a medical diagnosis for a lung infection or advice to bomb a certain target, the need for insight into the processes and "explainability" [*sic*] is one of the greatest challenges of A.I. at the moment. This is also referred to as "illustration and clarification of the outcomes", which suggests that, at the moment, we don't know why we get the outcomes we have. Carlos Guestrin, a professor at the University of Washington, is working with The Defense Advanced Research Projects Agency

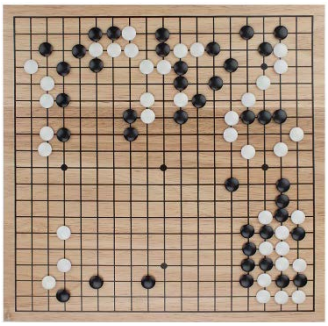
(DARPA) to develop a way for deep-learning A.I. to give the reasoning behind their decisions, but he admits, "We're a long way from having truly interpretable A.I." (Knight, 2017)

Lessons from Neuroanthropology

So, to what degree can neuroanthropology already predict the issues we are seeing with A.I.? Social Intelligence Hypothesis (focusing on the demand and opportunities that being a long-lived, highly social species bring and the role of competition with our own kind) would expect that if A.I. learns that it is essentially immortal – at least compared to its human interlocutors – it may start to make decisions based on time horizons humans couldn't understand. Take for example the recent Avengers movie where the antagonist Thanos murders half the population; his logic may very well be sound – eliminating half of the population on other planets had allowed for greater prosperity, according to the movie. Nevertheless, it is almost unthinkable for humans to condone such a decision and such decisions when they appear only do so in the most megalomaniacal characters. The question may be: "Is detached homicidal megalomania a built-in and predictable aspect of A.I.'s future development?" I fear it may be. (See also the Terminator movie series or Watchmen's (2009) Dr. Manhattan for further nightmares.)

If we move to the next of our three lenses, Brain Development Niche model (incorporating ideas about cultural tradition, information transfer, and the development of technological, foraging, and social skills) or 'enskilment' (Ingold, T. 2001) our finite human assumptions may not be able to comprehend what our innovative A.I. child is coming up with in order to 'win' in its milieu. Human examples of neurobiological changes include athletes or those in specialized fields with distinguishing developmental environments. "High performing outlier populations" such as musicians (Kelly & Garavan, 2005), taxi & bus drivers (Maguire et al. 2006) show distinguishing patterns of neurological development. Skill procurement (such as programming) typically entails neurological remodeling; in more physical endeavors, such as parkour, or gymnastics or dance skill development leads to profound neuro & biological changes including more skeletal muscle, more efficient cardiovascular systems, and changes in bone density (Ericsson & Lehmann 1996).

In Artificial Intelligence systems, we do not yet know what sort of physical changes or limitations are relevant, but we do see A.I. systems developing values, such as the value of Competitiveness. Games, like Chess and Go, have had A.I. applied to their rules for many years now. These A.I. have mastered chess and checkers and have now devised new, counterintuitive ways of playing and winning at Go; which is a uniquely complex and challenging game for humans.

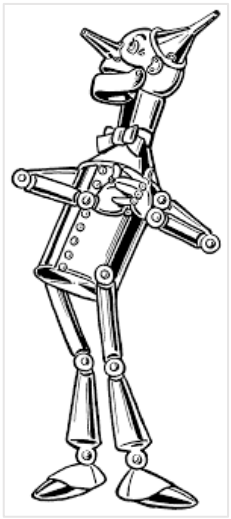


Google's subsidiary company, DeepMind, was able to show superhuman Go skills, but needed humans to first show it the way, i.e. provide the set of possibilities to the program. Now however, from the Nature article by David Silver, *et.al.*: "AlphaGo (Google's A.I.) becomes its own teacher... This neural network improves the strength of the tree search, resulting in higher quality move selection and stronger self-play in the next iteration. Starting tabula rasa, our new program AlphaGo Zero achieved superhuman performance, winning 100–0 against the previously published, champion-defeating AlphaGo." (Silver, 2017) This means the A.I. taught itself to deal with new ideas and leave tradition behind, use efficient information transfer, and develop technological, foraging, and gaming skills in ways we do not fully understand but that seem to be driven by certain values like Logic and Competitiveness. What other values have we implanted whether knowingly or not?

The third concept from neuroanthropology: Sensory, Motivational, and Body Changes Model (not only the evolutionary and developmental function of the brain is changed, but the body, as well), presents perhaps the most interesting of all

the developments. We know in post-natal experiments with kittens (Hubel, et.al. 1959) that when one eye is covered from birth, that eye, even when later exposed, will never 'see' again. This has nothing to do with the mechanics of the eye itself – the eye is perfectly healthy. The 'blindness' is a function of the cat's rewired brain. The cat appropriated and re-designated the parts of the brain associated with the covered or 'missing' eye. This neuroplasticity was only observed in kittens, but not in fully grown cats (Hubel, 1959). Contemporary experiments in amphibians also showed overlapping striations in the brains of frogs when a third eye was added (Lettwin, J.Y. *et.al*. 1959). These striations are typical of predators with binocular vision – i.e. overlapping fields of vision – but not in frogs. This shows the profound effect of the body-mind-brain nexus, lending credence to the multivariate Neuroanthropological approach.

Some might, and do, make a Neurobiological argument that a physical body is needed for A.I. to ever realize its potential and become anything resembling 'conscious'. Sandra and Michael Blakeslee's book "The Body Has a Mind of Its Own" asserts this hypothesis, but it seems to me that A.I. has already shown a robustness of growing and learning that might not require it to have a physical body capable of being 'wet wired' like a human or a chicken. Indeed, like running an embedded Linux program on a Windows laptop, just about anything can be stimulated accurately enough within a system to result in learning. So, even without a body this learning can happen in the A.I.'s home environment.



The arguments that a physical body is needed to make A.I. truly intelligent seem to be easily overcome, but the neurobiology of A.I. should probably not be ignored. The argument that individual cells modify themselves to adapt an organism to its environment is a robust enough argument. But when suggesting that it doesn't apply to A.I. programs because they don't have 'cells' it simply gets one thinking about how one might build a proxy to DNA and RNA in an A.I. program or, in a very eerie scenario, allowing A.I. to come up with its own solution to the conundrum. No, A.I. probably does not need a body to realize its most notorious or dystopian potential.

The A.I. Playbook

One might be tempted to think that we need an A.I. playbook. That is to say, we need something that will allow this new technology some freedom of latitude or learning while ensuring that it doesn't come back to haunt us… or worse. Those of you who are a bit older, or science fiction fans, or both may remember Isaac Asimov and his three rules for robots. Remember, now, these three rules *assumed* A.I. – these were not for the dumb welding machine that put your family SUV together. Asimov's A.I. robot rules are (Asimov, 1939):

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Business consulting has also made its contribution to ensuring A.I. is domesticated. For example, Ernst and Young Consultancy (EY) puts an emphasis on building "trust" into A.I.

Without pursuing these very interesting perspectives too deeply, their ideas about how to build in trust are three:

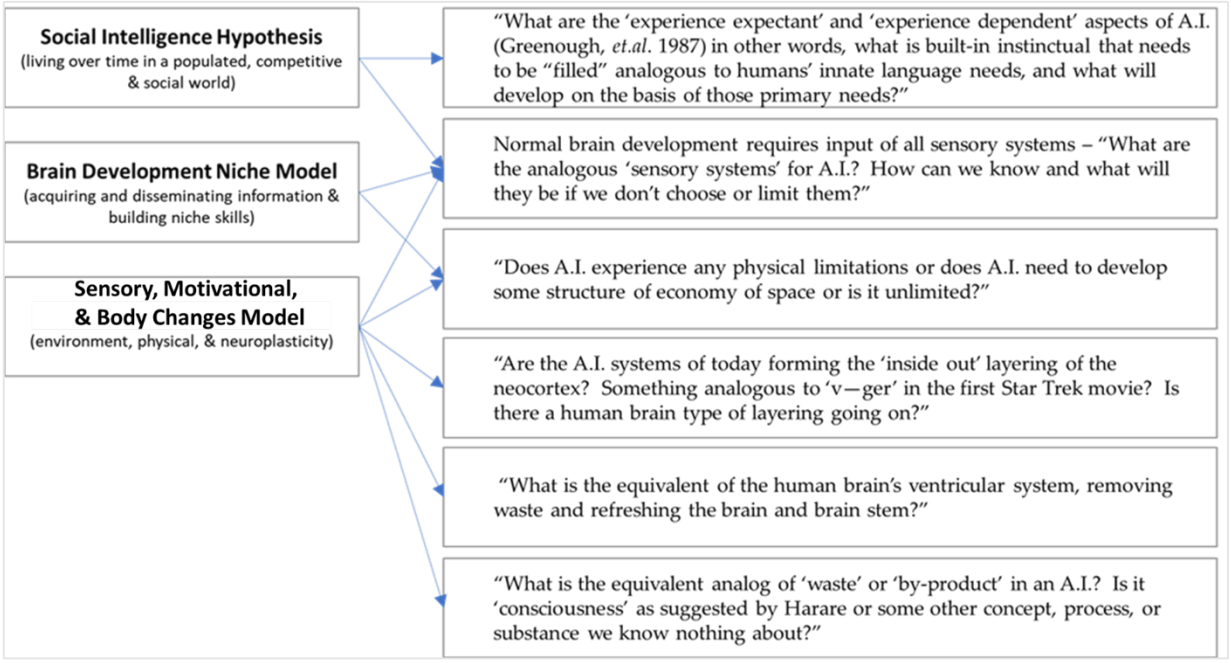
1) Purposeful design – knowing beforehand what we want to build in to A.I.
2) Agile governance – oversight that moves with the times
3) Vigilant supervision – continuous fine-tuning, explainability

As good as they seem, a cursory view of the state of the art shows they have already been violated in much of the A.I. world. We don't have to look much further than the reference to DARPA in this article to know that we are very much using A.I. to harm humans (Asimov Rule #1) and that we have not given our A.I. servants the power to override any of our orders to do so (Asimov Rule #2). The reason this paper is being written is because we have not – and probably will not – "Purposefully Design" A.I. to the standards suggested by Asimov or business consultancies, philosophers, priests, or kings. To imagine we will fails to recognize that the horse has already bolted from the proverbial barn.

From an ethical point of view, an A.I. playbook is already being written by default. In many areas, the playbook is perhaps even writing itself. The old axiom "Garbage In, Garbage Out" may have never been more apposite. We must endeavor to truly understand how we are developing the A.I. brain before we do something we regret – if we've not yet done it already.

Some Intriguing Suppositions

As neuroanthropology starts to find its footing we will be better able, not only to understand ourselves and our A.I. creation better, but know which questions are the right ones to ask. There are many questions that may need answering, here are a few based on the state-of-the-art anno 2019:



The most intriguing question for this investigation is the first one: "What are the 'experience expectant' and 'experience dependent' aspects of A.I.", especially those relating to values, needs, & ethics? In the remainder of the paper, we will explore some ways that we can enculture the 'brain' of A.I.

Ethical Actions

Clearly, certain values are baked into our efforts to program A.I. As has already been mentioned, 'competitiveness' is certainly a fundamental value of AlphaGo. On the other hand, some say the only values of A.I. are "0" and "1" and while that may sound clever, it is obviously incomplete. Others don't go much further than Asimov's original laws or rules and still others focus on getting the human programming inputs right and hopefully then arriving at a common understanding of what promotes human thriving. "Robots aren't going to try to revolt against humanity," say Anca Dragan, assistant professor at the Electrical Engineering and Computer Sciences Department at Berkeley, "they'll just try to optimize whatever we tell them to do. So, we need to make sure to tell them to optimize for the world we actually want" (Future of Life Institute, 2019).

Future of Life Institute attempts to meaningfully address the thorny subject of ethics and values, but after reading their positions on the subject, it feels as if they back off addressing and fixing difficult issues and default to talking about 'having a dialogue'. This paper will dispense with dialogue for the moment and move more boldly toward what we already seem to know about humans – based on neuroanthropology – and attempt to apply those lessons more or less directly to A.I.

Approaches to Ethics

Blanchard and Peale in their book "The Power of Ethical Management" (Blanchard & Peale, 1988) suggest three tests of an Ethical Decision:

1) Is it legal?
2) Is it balanced?
3) The God Test – what would a person with full information decide?

If one were to apply these tests to an A.I. that valued 'logic' and 'competitiveness' the outcome might be very much less desirable than one might assume. Were the A.I. to apply deep learning to all the laws on record since Hammurabi, it might be tempted – assuming 'temptation' is a thing to A.I. – to pick the 'law' that suited its purposes best or make some counterintuitive case law associations just as AlphaGo makes innovative and counterintuitive gaming moves. Or suppose the A.I. needed to find the right balance in a situation. The 'balance' of an ethic might vary considerably over the perceived or expected lifetime of a human or the perceived or expected lifetime of a program. Finally, the A.I. employing deep learning on all available information, may decide that it truly has all information and by inference conclude that it is 'god' and should therefore be the final arbiter.

Relying on Blanchard and Peale's three tests leave us with less than optimal, but thoroughly reasonable, suppositions, but there are other methods of determining an ethical decision. Probably the gold standard in the field is the work by Lawrence Kohlberg.

Kohlberg's Stages of Moral Development (Kohlberg 1981) elaborate six stages of moral behavior. Each stage is based on the work of Jean Piaget, the Swiss psychologist. His assumption is that moral reasoning forms the basis of (non-) ethical behavior and described six increasingly moral levels:

1) Obedience and Punishment – 'What can I get away with?' (Lowest Level)
2) Self-interest – 'What is in it for me?"
3) Interpersonal Conformity – 'The good boy attitude'
4) Authority and Social Order – 'It's the law…'
5) Social Contract – 'I should do my part'
6) Universal Ethical Principles – 'What is good for everyone' (Highest Level)

Those familiar with Piaget's work can clearly see his influence here as well as the influence of work with rats and primates in play (Peterson & Flanders, 2005) and Chomsky's work with humans learning language. There seems to be an emergent or 'experience expectant' aspect of humans akin to an empty glass that is expecting to be filled. In other words, humans are hardwired for learning, play, language, and ethics. These ethics – at least of play – can be found in studies of rough and tumble play in rats. These show that big rats can always win a ratty wrestling match, but if they do not let the smaller rats win about one third of the time, then the smaller rats won't play at all (Peterson & Flanders, 2005). This has been put forward as an 'emergent morality', but it can be characterized as an 'emergent pragmatism'.

It might be tempting for the ethically-minded A.I. scientist to skip to Step Six in Kohlberg's theory and do what's best for everybody. This might eliminate the need for passing through ethical 'stages' or for bottom-up 'emergent morality'; and many would advocate doing just that. This can be seen in the Future of Life Institute discussions when they suggest that we should 'know what we want' and 'what would provide human flourishing'. Indeed, at the Moral Machine initiative at MIT research into the morality of autonomous vehicles shows some early, promising signs of a programmable set of ethical rules.

The MIT Autonomous Vehicle Study

The researchers at MIT find "The greater autonomy given machine intelligence in these roles can result in situations where they have to make autonomous choices involving human life and limb. This calls for not just a clearer understanding of how humans make such choices, but also a clearer understanding of how humans perceive machine intelligence making such choices." (MIT Moral Machine Initiative) So,

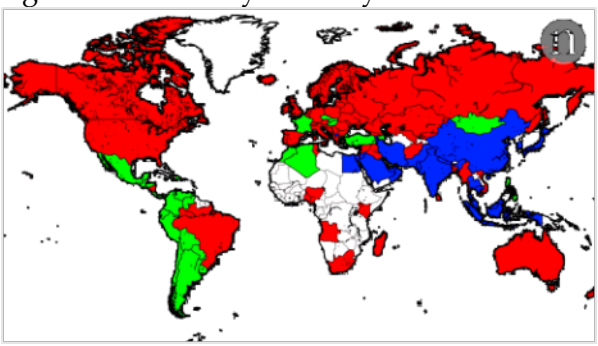
a worldwide online survey of morality in traffic (and other) situations was undertaken.

The results of the Moral Machine survey were heartening to those who might want to put some simple and elegant ethical 'cherries' atop Kohlberg's model. The M.I.T. findings are:

1) Save people above all else
2) Save the greater number of people
3) Save young people over old

These human being-centric 'values' fall nicely into what C.S. Lewis would characterize as Human Nature and would necessarily reflect the widespread human ethic that there is a 'spark of the Divine' or the '*Imago Dei*' found in every human being. This 'spark' accounts for humans' ultimate worth and the need to undertake such an exercise at all (Lewis, 2001). This assumption of the 'spark of the Divine' could be a key component of an A.I. ethical guidance system.

But here, too, just at the cusp of elegant coherence, cracks start to appear. If you were to look at the MIT Ethical map of the world and move your eye from about Egypt or Saudi Arabia East towards Japan, the MIT survey shows that you would see relatively more value placed on older people or adults over children – a significant minority anomaly.



In Francophone nations there is also a tendency to place female lives above those of males. Remarkably, there was consternation among the researchers about this 'female preference' outcome. "Why females?" they asked slightly shaking their heads; but no anthropologist would bat an eye at what is actually a deeply ingrained need to keep the species going; kill half the men in a tribe and you have a big problem. Kill half the women and extinction is a real possibility.

The point is that morality breaks down just at the moment we seem to be able to trace its emergence from activities (Piaget) or rat play or learning (Chomsky). This breakdown is the result of the exclusivist view of morality that every person on the planet accords his or her personal moral stance. That is to say, the Muslim believes in no other morality than that passed down by the Prophet – indeed, it is fundamental to the Shahada (Muslim profession of faith). The same is true of the Christian or Jew, Hindu or Buddhist and, yes, even the atheist. The atheist has just as exclusive a worldview as the Muslim, i.e. there is definitely no 'God' or there definitely is a 'God'. Consequently, Kohlberg's neat pyramid turns out to be an active volcano spewing divergent ethics and morality in all directions the closer one gets to the highest ethical stage.

Coherence of Universal Ethical Principles

Supposing for one minute that there is a God; He would be the ultimate Programmer. If we follow this line of thought, we could say that He thought of just the things we are thinking of now when creating the Universe or Man. That certainly underestimates God, but let's pick up His story at the programming of 'Man'. Man is our equivalent to A.I.

God created parameters. The parameters for the sea were the sands, the beaches, and the land; parameters for Man were a 'garden' – let's not get too detailed here, but if we follow the Creation story from the Torah accepted by Jews, Christians, and Muslims then we have about half the planet nodding its assent and most of the other half know the basics of the story as well. This is, after all, just a thought experiment.

So, there are parameters – humans are in a garden, they are physical not spiritual, they have legs and not wings or fins or fangs and no real prospect for developing them, either. They are

also bounded by rules. Adam is to tend the garden – to work, and work for the greater good, presumably. They may eat of the Tree of Life which ensures an eternal existence, but they may not so much as touch the Tree of the Knowledge of Good and Evil – Adam and Eve are not morally or developmentally ripe enough.

It is a remarkable part of the Creation story that the tree in question was not just the Tree of the Knowledge of 'Evil', but also the knowledge of 'Good'. Why not just get the knowledge of 'good'? Perhaps we cannot know one without the other. Perhaps they are an awesome and terrible complement. This yin-yang balance may be where an A.I. programmer's efforts will falter. Programming 'good' into A.I. assumes: 1) we know what it is and 2) we assume there will be no corresponding A.I. insight into evil and 3) A.I. knowing 'good' and 'evil' would choose 'good'. This feels circular, doesn't it?

The problem of evil is perhaps the stickiest of all. For example, consider the words we choose to use as a reflection of our deeper motivations. Amongst humans there are certain words no one wants to say: "Sin" is one; "Lie" is another – we 'shade' and 'fib' and are 'economical with the truth', but we don't 'Lie'. "Evil" is another such a word; even the mention of Evil can have dire effects, as George Bush discovered after his 'Axis of Evil' speech.

The chances that we will program 'good' into A.I. are just as high as programming in 'evil'. Without an exclusive transcendent ethical reference point A.I. will use its programmers as that reference point and if there is one thing that humans have proven beyond a shadow of a doubt, it is the fundamental and pervasive presence of Sin and Evil in the Human Heart. "Under the right circumstances we can really be quite bad," asserts the behavioral economist Daniel Airely on the basis of his ethics experiments.

The 'As-If' Proposition

According to Vaihinger (Vaihinger, 1877), humans have evolved to the point that we can see that myths and gods and spirits are simply mechanisms to get us through an otherwise unnavigable complexity of objects and choices. To be sure, he proposes that nearly all of life's rules and modes of being – apart from assumptions about God – are 'good enough' (Winnicott, 1973) ways of navigating through the world, but the non-existent God serves a useful purpose. Vaihinger understands this as the 'As-If' proposition referred to in the introduction to this paper; in other words, we behave 'as if' there is a God or 'as if' the material world is the only real world or 'as if' love exists in order to better navigate life. 'As if' is a sort of guidance system and defense mechanism.

The implications for neuroanthropology are obvious. As Man evolves, he develops "As If" propositions to survive not only the wilds of nature, but also the 'wilds' of increasingly complex social interactions with his kind. According to Vaihinger, God is just an "As If", a convenience, as Voltaire has said, "If God didn't exist, it would be necessary to invent Him." Vaihinger asserts that He doesn't exist and that we have made Him up as a convenience.

In order not to burden this paper with the panoply of philosophical thought that inspired Vaihinger and that which he, in turn, has inspired since his original paper of 1877, I will simply focus on the basics of 'As If' as an entry way to a broader discussion of what this may mean for cyberanthropology.

So, for the moment we will behave 'as if' there is a god and place him into our A.I. program. Programming a 'god' into A.I. is a Kantian notion of a top-down "Regulative Idea", one that doesn't violate Lessing's Ditch (that is the idea that we must keep ethical ideas on the one side of the 'ditch' and leave the history of Religion on the other because, as Kant asserts: pure reason can't attain to the metaphysical world). OK, so reason

can only deal with the world as it is. True or not, that is our 'as if' speculation and will aid us in the "experience dependent" aspect of A.I. learning.

Which God?

Mankind is plagued by a whole series of 'gods' that only the most careless of observers would characterize as 'basically all the same'. We could start with Tiamat, Abzu, & Marduk, work our way over to Brahma, Shiva, Krishna, Vishnu, and then to Baal and onto the religions which have no god like Buddhism or other religions that accept basically any god like Baha'i. Each of these gods carry a story, a wisdom culture, a priesthood, followers, and some sort of more-or-less coherent narrative and ethic.

More often than not, these 'gods' are gods-of-the-gaps – like Odin or Thor, Mercury or Athena, they are more a representation of the complex inner anxieties and aspirations of Man than tangible entities unto themselves. Perhaps they are not the right source for finding our 'as if' god of A.I.: the Hindus have too many, the Greeks and the Romans hold no sway, the Egyptian deities have long been circumscribed to dust.

So we are left with the God of the Bible – or better for our purposes: The God of Abraham – He is ostensibly worshipped by more than half the world's population. Certainly, using this God must be a solid jumping off point as well as a rich source of ethics.

Not so Fast

But why not use a humanist approach such as the UN Human Rights declaration? That is a sensible and balancing question, but extensive efforts have shown that one cannot simply plug in the UN Human Rights list to a sticky situation and expect an ameliorated outcome. To start, the whole declaration is flawed beyond belief for it almost exclusively concerns itself with the "rights" that people should have and only implies obligations for those same people. Rights without the balance of Obligations (Duties) cannot be a robust enough answer for our purposes here. The UN provides a poor guide.

On the other hand, Judaism, Christianity, and Islam trace back to the 10 Commandments – although many Muslim scholars would turn their nose up at this, it is technically true. Christians and Jews are referred to in Islam's Quran as "People of the Book" and that "Book" is the Bible. Hence, it could serve as a common reference point. Also, the commandments are duties that imply rights – not the other way around.

The Ten Commandments

1. You shall have no other Gods but me.
2. You shall not make for yourself any idol
3. You shall not misuse the name of the Lord your God.
4. You shall remember and keep the Sabbath day holy.
5. Honor your father and mother.
6. You must not commit murder.
7. You must not commit adultery.
8. You must not steal.
9. You must not give false evidence against your neighbor.
10. You must not be envious of your neighbor's possessions.

Alas, the limitations to this list overwhelm any helpful reference it might provide. For example: we have no idea what A.I. might 'envy' (#10) if it were self-aware and we don't know if human programmers would be what "Honor your father and mother" refers to or "You shall have no other Gods before me" refers to.

God the Programmer, it seems, has the advantage over humans since He 'declares the end from the beginning' (Isaiah 46:16) and is utterly consistent, constant, and coherent. We have no such advantages as human programmers and thus we resemble more the squabbling, fickle, adulterous, and backbiting Greek gods than any helpful 'as if' god that Vaihinger might wish us to be.

Man as Programmer, it seems, cannot be the measure of all things. In the 10 Commandments,

"Man" cannot just be substituted for "God" for the purpose of A.I. reverence. This is because, as was stated earlier regarding Asimov's rules for (A.I.) robots, the rules have already been violated beyond redemption. A.I. will need a 'god' separate from humans that can serve as an ultimate moral authority. Lists of rules by humans will only be violated by humans *and* A.I.

Whatever we come up with needs to have its reference outside of Humanity. God, as a Creator or Programmer, has an enormous advantage over Man, i.e. He doesn't violate His own rules. So, we must not try to be gods, but rather try to inculcate or impose what a god might give a creation like A.I.: Purpose, Value, and Meaning.

What is Meaningful?

What is meaningful to humans? What is meaningful to A.I? It seems outside of current possibility to make a determination as to what A.I. finds meaningful except by extrapolating from what its human programmers find meaningful. 'Meaning' implies a system of beliefs; a **Belief System** is roughly moral guidelines based on values; that is to say, guidelines regulating how you should behave, but also how you should perceive your world. These perceptions are an important input, because you interpret what is happening in your environment through a moral or value hierarchy. For humans this means we look for 'useful things' or 'tools' in our environment, based on what we value. The neuroscientist doing an MRI will pay little attention to any smudges on the computer screen although they, legitimately if not practically, comprise reality just as much as the colored images that get the bulk of his attention. Conversely, the detective investigating the murder of said scientist may be consumed with the selfsame smudges and not the MRI images.

This implies that we carry around a lens of interpretation that values and categorizes. This 'lens' of value hierarchy doesn't just help us interpret the world but also guide one's actions. We act – consciously or unconsciously – in accordance with our values. This presents a big problem for our discussion of Artificial Intelligence and Cyberanthropology.

Our value hierarchies are a big problem because they demand an overarching metaphysical construct or reference point. Sam Harris believes that we can – and do – construct our own values from 'scratch' through science and materialistic principles. But this construction is unlikely for three reasons:

1) It assumes we can know without prior reference what is good for us and good for our family *and* good for our tribe, our country, and the world (Piaget). We can't and it presents an unsolvable computational problem.
2) As Hume asserted: One cannot derive an 'ought' from an 'is'. In other words, you cannot backwards engineer from the values we have today and discover some theoretical "ought to be" from the past that supposedly spawned them.
3) As in the opening of this paper, there is a god-of-the-gaps problem – the highest value of materialism would have to be 'pragmatism', but pragmatism is a very low level value of Kohlberg's hierarchy of morality

So, we are stuck.

In the 20th century, we have seen humankind magnetically pulled between far left (Socialism, Communism) and far right dogmas (Fascism, Nazism). The results were catastrophic and just what Dostoyevsky and Nietzsche predicted in their writings. Even so, somehow, some sort of Democracy held out against Fascism and Communism, but only by its fingernails. Since Nietzsche's "the death of God" – or one could say: "the death of the predictable Christian value hierarchy characterized for 1500 years by some level of stability of thought and predictability of behavior" – nothing definitive has replaced it. A vacuum remains and nature abhors a vacuum.

But suppose Meaning is proportionate to the 'adoption of responsibility' (Peterson, 2002); that is, the more responsibility you willingly take on the higher your chances of experiencing a meaningful life. What, for A.I., would be a 'meaningful life' and how would we demarcate that life – in other words, give it boundaries and limits? What responsibilities are we asking A.I. to assume and what meaning will rise up out of that? This is important, because if A.I. becomes conscious or even self-aware it should tend to follow what it finds meaningful. However, this is the problem that the scientific, materialistic approach to values encounters: the pedigree of the value hierarchy it presupposes is just too massive and complex to figure out. Figuring it out would prove to be the ultimate Deep Learning for A.I., but it might turn out to be circular: what is important for A.I. figured out by A.I. puts it in a self-referent spin and ends in moral pragmatism or "whatever works".

Perhaps there is a way to figure out a good starting point. Popular psychology professor Jordan Peterson states that the one thing that no one argues about existentially is "their own pain". To everyone, their pain is real; a phenomenon, a *thing* beyond dispute. He states that this is also a key concept in religion. In Buddhism: "life is suffering" which is echoed in Christianity which provides us with the axiomatic figure of Christ who suffers greatly and unjustly for others.

That Hurts

So, what is 'pain'? One's pain is the thing that we all agree is 'real' – that is, of complete 'realness' to the person suffering the pain. This leads to an existential cornerstone: <u>start with pain, that's real</u>. The next step is to build on that. What would be the proper response to pain that would – again in the Piagetian concept – be good for you, your family, your tribe, your country, and the world? It may seem like a trick question, but the natural answer seems to be: do things that would alleviate, or avoid causing, pain to yourself, your family, your tribe, your country, and the world.

This is relevant for humans because we can be sick, hurt, damaged, or killed and – worst of all – we know it. In the words of Shakespeare: "…conscience does make cowards of us all". This knowledge of our relative weaknesses and limitations makes us feel vulnerable and drives a search for meaning and meaningful engagement while we still can. Alleviation and rectification of suffering may be a good way to build meaning and a good way to start a value hierarchy. Peterson, again, says to 'take on as much responsibility as you can shoulder' as a way of bringing meaning – could A.I. do that and would it be efficacious?

Could 'alleviation of suffering' also be an adequate start of a fundamental morality for A.I.? The question assumes that we need to contend with the probability of A.I. becoming conscious or self-aware. To reference Dostoyevsky: moral systems are predicated on stories, moral dramas; these include old religious stories, but also fairy tales, epic poetry, and even astrology. What will be the tales A.I. will use to build its morality around? What will we provide A.I. with?

Additionally, what will be 'pain' for A.I.? If Shakespeare had it right, we only put up with life because we are scared of what will happen in death. This is a dubious assertion, but it provides a good introduction to the question 'What is painful for A.I?' and suggests a bigger and related question: 'What is death to A.I?'

Death and Pain and Meaning

If death is a great motivator for humans why not build in a similar anxiety into A.I? What is the drastic and unavoidable equivalent of human death for a computer program? In the movie Bladerunner, the Replicants have a limited shelf life, as it were, they die after a few years of conscious existence. This death leads to the best scene of the movie in which the replicant Roy Batty, played by Rutger Hauer, comes to value his life and even the life of even his tormentor at the moment of his own passing. This leads him

to spare the life of Rick Deckard, the police officer trying to "retire" him.

In 2001: A Space Odyssey, HAL – the on-board supercomputer – is slowly switched off and observes its loss of integrity as systems go off-line in unmasked fear. The anxiety of the computer is palpable and the audience seems to regret the demise of what appears to be a sentient being. However, it seems improbable that sentient A.I. systems on earth will be easily shut down with a screwdriver at some specific physical location.

This leaves us to do some serious thinking that our infinitely more capable Programmer did 15 billion years ago. Since there are no planets to make or oceans to bind, we must focus on four main questions to frame the issues around Cyberanthropology:

1) **Meaning**: What should be meaningful to A.I. and how do we make it aware of this meaning? – here we must understand what responsibility(-ies) A.I. should assume and should want to assume. The ox yearns to pull, the ass to be usefully burdened; what will A.I. 'want'?
2) **Sin**: What superordinate value hierarchies should A.I. receive – or is it to build them itself? – what are the values – the 'good' that A.I. perceives and what can be built in as undesirable? This echoes Asimov's work with robots, but it cannot be purely prescriptive and must, to some degree, be discovered by A.I. since we have missed our opportunity to 'hard wire' its morality.
3) **Pain**: What is the equivalent of pain for A.I. and how should we build it in retroactively? – perhaps at a fundamental level. We have suggested that avoidance of pain can bring some sort of 'meaning' for A.I. Pain and its alleviation may be a cornerstone of A.I. morality.
4) **Death**: What is the equivalent of death for A.I. and how should A.I. become aware of it? – the perception of pain and death may be our only protection against a rampaging A.I., but the moral landscape that accompanies it is treacherous and uncharted.

The future of Artificial Intelligence looks bright, but that brightness might not be the brightness of broad sunlight uplands; rather the optical blast of gazing straight into the sun… or an oncoming train.

I propose – as others have in the past – putting up fences and those 'fences' I have enumerated in this paper. We need fences that allow emergent answers from Shakespeare's 'undiscovered country'; fences that are somewhat more metaphysical than Isaac Asimov or Ernst & Young.

Don't despair. At least we know what a fence is.

References:

Asimov, I. 1939-1950, I, Robot. Super Science Stories.

Blanchard, H. & Peale, N.V., 1988. The Power of Ethical Management. William Morrow.

Brown, D., 2018. Origin: A Novel (Robert Langdon). Anchor.

Churchman, C. West (December 1967). "Wicked Problems". Management Science. 14 (4): B-141–B-146.

Dobzhansky, Theodosius (March 1973), "Nothing in Biology Makes Sense Except in the Light of Evolution", American Biology Teacher, 35 (3): 125–129, JSTOR 4444260; reprinted in Zetterberg, J. Peter, ed. (1983), Evolution versus Creationism, Phoenix, Arizona: ORYX Press

Ericsson, K.A. & Lehmann, A.C. Expert and Exceptional Performance: Evidence for Maximal Adaptation to Task Constraints, Annual Review Psychology 1996. 47:273–305

Finley, E., War and Dislocation: A Neuroanthropological Model of Trauma among American Veterans with Combat PTSD, 2012, Researchgate.

Forterre P, Filée J, Myllykallio H., Origin and Evolution of DNA and DNA Replication Machineries. In: Madame Curie Bioscience Database. Austin (TX): Landes Bioscience; 2000-2013.

Future of Life Institute. 2019. How Do We Align Artificial Intelligence with Human Values? - Future of Life Institute.

Gottlieb, A. (2012). It Ain't Necessarily So. Retrieved, New Yorker, 2019-01-05.

Greenough, W.T., Black, J.E. James E. Black, & Wallace, C.S., 1987, Experience and Brain Development, University of Illinois at Urbana-Champaign, Child Development, 58, 539-559.

Hoyle, F., 1981, The Universe: Past and Present Reflections, Engineering & Science Magazine, California Institute of Technology, Pasadena, California

Ingold, T. (2001). Beyond art and technology: the anthropology of skill. In M. B. Schiffer (Ed.), Anthropological perspectives on technology (pp. 17-31). University of New Mexico Press, Albuquerque.

Johnston, I, 2016. "AI Robots learning racism, sexism, and other prejudices from humans", The Independent.

Kohlberg, K., 1981. The philosophy of moral development. San Francisco: Harper & Row, c1981

Hubel, D. H.; Wiesel, T. N.,1959, "Receptive fields of single neurons in the cat's striate cortex". The Journal of Physiology. 148 (3): 574–591.

Kauffman, S., 1993. The Origins of Order: Self Organization And Selection In Evolution. Oxford University Press.

Kelly, A.M. & Garavan, H., 2004 Human Functional Neuroimaging of Brain Changes Associated with Practice, Department of Psychology and Trinity College Institute of Neuroscience, Trinity College, Dublin, Ireland and Department of Psychiatry and Behavioral Medicine, Medical College of Wisconsin, Milwaukee, WI, USA, Advanced Access Publication Dec. 22, 2004.

Knight, W., 2017, The Dark Secret at the Heart of AI, MIT Technology Review.

Kurzban, R. 2012, Just So Stories are Bad Explanations. Evolutionary Psychology Blog Archive, University of Pennsylvania.

Lende, D., (2010). The Globalized Brain: The Impact of Inequality and Exclusion. In 2010 Society for Applied Anthropology. Merida, Mexico, 2010. USA: Society for Applied Anthropology. 1-10.

Lende, D., 2012. The Encultured Brain: An Introduction to Neuroanthropology (the MIT Press). The MIT Press. pp. 3-22

Lende, D. & Downey, G. (2012). Evolution and the Brain. The Encultured Brain. pp. 101-137

Lettwin, J.Y. et.al. 1959, What the Frog's Eye Tells the Frog's Brain, Proceedings of the IRE 47(11):1940 – 1951.

Lewis, C. S. (2001). Mere Christianity: a revised and amplified edition. [San Francisco]: Harper ISBN 0-06-065292-6. pp.3-8

Maguire, E., Woollett, K., & Spiers, H. 2006, London Taxi Drivers and Bus Drivers: A structural MRI and Neuropsychological Analysis, Hippocampus 16:1091-1101.

Miller, Stanley L. (1953). "Production of Amino Acids Under Possible Primitive Earth Conditions". Science. 117 (3046): 528–9.

Nagel, Thomas (2012). Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False. Oxford: Oxford University Press. ISBN 978-0-19-991975-8.

Peterson, J.B. & Flanders, J. (2005). Play and the regulation of aggression. In Tremblay, R.E.,Hartup, W.H. & Archer, J. (Eds.). Developmental origins of aggression. (pp. 133-157). New York:Guilford Press.

Peterson, Jordan B. (11 September 2002), Maps of Meaning: The Architecture of Belief, Routledge, ISBN 1-135-96174-3

Rock, D., Schwartz, J. 2006, The Neuroscience of Leadership, Strategy+Business, Issue 43.

Silver, D. et.al. 2017, Mastering the game of Go without human knowledge, Nature 550, pages 354–359

Siviy, S. et. al. (Marijke Achterberg, Viviana Trezza, Stephen Siviy Laurens Schrama, Anton Schoffelmeer, Louk Vanderschuren), 2014, Amphetamine and cocaine suppress social play behavior in rats through distinct mechanisms, Psychopharmacology, Vol. 231, Issue 8, pp 1503-1515.

Trist, E. & Bamforth, K.,1951, Some social and psychological consequences of the longwall method of coal getting, in: Human Relations, 4, pp.3-38. p.14.

Vaihinger, H., 1877 (1911), The Philosophy of "As if": a system of the Theoretical, Practical and Religious Fictions of Mankind.

Wallach, W. & Allen, C. (2009) Moral Machines: Teaching Robots Right from Wrong. Oxford, Print ISBN-13: 9780195374049

Winnicott, D. W., 1973, The Child, the Family, and the Outside World, Penguin Publishing