Roman V. Yampolskiy

Artificial Consciousness: An Illusionary Solution to the Hard Problem

(doi: 10.12832/92302)

Reti, saperi, linguaggi (ISSN 2279-7777)
Fascicolo 2, luglio-dicembre 2018

ARTIFICIAL CONSCIOUSNESS
AN ILLUSIONARY SOLUTION TO THE HARD PROBLEM

Roman V. Yampolskiy

The Hard Problem of consciousness has been dismissed as an illusion. By showing that computers are capable of experiencing, we show that they are at least rudimentarily conscious with potential to eventually reach superconsciousness. The main contribution of the paper is a test for confirming certain subjective experiences in a tested agent. We follow with analysis of benefits and problems with conscious machines and implications of such capability on future of computing, machine rights and artificial intelligence safety.

Keywords: Artificial Consciousness, Illusion, Feeling, Hard Problem, Mind Crime, Qualia.

> «The greatest obstacle to discovery is not ignorance – it is the illusion of knowledge».
>
> Daniel J. Boorstin

> «Consciousness is the one thing in this universe that cannot be an illusion».
>
> Sam Harris

1. INTRODUCTION TO THE PROBLEM OF CONSCIOUSNESS

One of the deepest and most interesting questions ever considered is the nature of consciousness. An explanation for what consciousness is, how it is produced, how to measure it or at least detect it (Raoult, Yampolskiy 2015) would help us to understand who we are, how we perceive the universe and other beings in it, and maybe even comprehend the meaning of life. As we embark on the quest to create intelligent machines, the importance of understanding consciousness takes on the additional fundamental role and engineering thoroughness. As presence of consciousness is taken to be the primary reason for gran-

ISSN 1826-8889

ting many rights and ethical consideration (Muehlhauser 2017), its full understanding will drastically change how we treat our mind children and perhaps how they treat us.

Initially the question of consciousness was broad and ill-defined encompassing problems related to intelligence, information processing, free will, self-awareness, essence of life and many others. With better understanding of brain architecture and progress in artificial intelligence and cognitive science many easy sub-problems of consciousness have been successfully addressed (Chalmers 1995) and multiple neural correlates of consciousness identified (Mormann, Koch 2007). However, some fundamental questions remain as poignant as ever: What is it like to be bat? (Nagel 1974), What is it like to be a brain simulation? (Özkural 2012), etc. In other words, what is it like to be a particular type of an agent (Burn 2008; Laureys, Boly 2007; Preuss 2004; Trevarthen 2011)? What it feels like to be one? Why do we feel something at all? Why red doesn't sound like a bell (O'Regan 2011)? What red looks like (Jackson 1986)? What is it like to see with your tongue (Kendrick 2009)? In other words, we are talking about experiencing what it is like to be in a particular state. Block (Block 1995) calls it *Phenomenal* or P-consciousness to distinguish it from *Access* or A-consciousness. David Chalmers managed to distill away non-essential components of consciousness and suggested that explaining qualia (what it feels like to experience something) and why we feel in the first place as opposed to being philosophical zombies (Chalmers 1993) is the Hard Problem of consciousness (Chalmers 1995):

> The really hard problem of consciousness is the problem of *experience.* When we think and perceive, there is a whir of information processing, but there is also a subjective aspect. As Nagel (1974) has put it, there is *something it is like* to be a conscious organism. This subjective aspect is experience. When we see, for example, we *experience* visual sensations: the felt quality of redness, the experience of dark and light, the quality of depth in a visual field. Other experiences go along with perception in different modalities: the sound of a clarinet, the smell of mothballs. Then there are bodily sensations from pains to orgasms; mental images that are conjured up internally; the felt quality of emotion; and the experience of a stream of conscious thought. What unites all of these states is that there is something it is like to be in them. All of them are states of experience (Chalmers 1995).
>
> […A]n organism is conscious if there is something it is like to be that organism, and a mental state is conscious if there is something it is like to be

> in that state. Sometimes terms such as «phenomenal consciousness» and «qualia» are also used here, but I find it more natural to speak of «conscious experience» or simply «experience» (Chalmers 1995).

Daniel Dennett (Dennett 2017) and others (Tye 1999) have argued that in fact there is no Hard Problem and that what we perceive as consciousness is just an illusion like many others, an explanation explored by scholars of illusionism (Balog 2016; Blackmore 2016; Frankish 2016; Tartaglia 2016). Over the years a significant amount of evidence has been collected all affirming that much of what we experience is not real (Noë 2002), including visual (Changizi *et al.* 2008; Coren, Girgus 1978; Gregory 1997), auditory (Deutsch, 1974), tactile (Nakatani Howe, Tachi 2006), gustational (Todrank, Bartoshuk 1991), olfactory (Herz, von Clef 2001), culture specific (Segall *et al.* 1963) and many other types of illusions (Kahneman, Tversky 1996). An illusion is a discrepancy between agent's awareness and some stimulus (Reynolds 1988). Illusions can be defined as stimuli which produce a surprising percept in the experiencing agent (Bertamini 2017) or as a difference between perception and reality (Zeman 2015). As we make our case mostly by relying on Visual Illusions in this paper, we include the following definition from García-Garibay *et al.*: «Visual illusions are sensory percepts that can't be explained completely from the observed image but that arise from the internal workings of the visual system.» (García-Garibay, de Lafuente 2015).

Overall, examples of illusions may include: impossible objects (Penrose, Penrose 1958), blind spot (Tong, Engel 2001), paradoxes (Zeno's – Misra, Sudarshan 1977 –, mathematical/logical illusions – Grelling 1936 –), quantum illusions (Greenleaf *et al.* 2011), mirages (Luckiesh 1922), art (Escher 2000; Gold 1993), Rorschach tests (Lord 1950), acquired taste (Mennell 1996), reading jumbled letters (Velan, Frost 2007), forced perspective (Kelley, Endler 2012), gestaltism (Koffka 2013), priming (Tulving, Schacter 1990), stereograms (Becker, Hinton 1992), delusion boxes (Ring, Orseau 2011), temporal illusions (Eagleman 2008), constellations (Liebe 1993), illusion within an illusion (Deręgowski 2015), world (Bostrom 2003), déjà vu *(Bancaud et al.* 1994), reversing goggles (Wallaeh, Kravitz 1965), rainbows (Fineman 2012), virtual worlds (Rheingold 1991), and wireheading (Yampolskiy 2014). It seems that illusions are not exceptions, they are the norm in our world, an idea which was rediscovered through the ages (Gillespie 2006; Plato, Grube 1974; Sun 1924).

Moreover, if we take a broader definition and include experiences of different states of consciousness, we can add: dreams (including lucid dreams (Barrett 1992) and nightmares (Zadra, Donderi 2000)), hallucinations (Bentall 1990), delusions (Garety, Hemsley 1997), drug induced states (Becker 1967), phantom pains (Carlen *et al.* 1978), religious experiences (Fenwick 1996), self (Hood 2012) (homunculus, Dennett 1981), cognitive biases (Gigerenzer 1991), mental disorders, invisible disabilities and perception variations – dissociative identity disorder (Kluft 1996), schizophrenia (Dima *et al.* 2009; Keane *et al.* 2013), synesthesia (Cytowic 2002), simultanagnosia (Coslett, Saffran 1991), autism (Happé 1996), ideasthesia (Jürgens, Nikolić 2014), asperger's (Ropar, Mitchell 1999), apophenia (Fyfe *et al.* 2008), aphantasia (Zeman *et al.* 2015), prosopagnosia (Damasio *et al.* 1982) – all could be reclassified as issues with «correctly» experiencing illusions), pareidolia (Liu *et al.* 2014), ironic processes (Wegner 1994), emotions (love, hate) (Izard 1991), feelings (hunger, pain, pleasure) (Harlow, Stagner 1932), body transfer (Slater *et al.* 2010), out of body experiences (Ehrsson 2007), sensory substitution (Bach-y-Rita, Kercel 2003), novel senses (Gray 2000), and many others.

Differences between what is traditionally considered to be an illusion and what we included can be explained by how frequently we experience them. For example, the sky looks different depending on the time of day, amount of Sun or the angle you are experiencing it from, but we don't consider it to be an illusion because we experience it so frequently. Essentially, everything can be considered to be an illusion, the difference is that some stimuli are very common while others are completely novel to us, like a piece of great art, see for example (Escher 2000). This makes us think that if we experience something many times it is real, but if we see something for the first time it must be an illusion.

At the extreme, we can treat every experience as an illusion in which some state of atomic particles in the universe is perceived as either a blue sky, or a beautiful poem or a hot plate or a conscious agent. This realization is particularly obvious in the case of digital computers, which are machines capable of extrapolating all the world's objects from strings of binary digits. Isn't experiencing a face in a bunch of zeroes and ones a great illusion, in particular while another machine experiences a melody on the same set of inputs (Wells 2012, 44)?

Likewise, neurodiverse individuals may experience the world in very different ways, just consider color blindness (Post 1962) as example of same inputs being experienced differently by diverse types of human agents. In fact, we suggest that most mental disorders can be better understood as problems with certain aspects of generating, sustaining or analyzing illusions (Dima *et al.* 2009). Similarly, with animals, studies show that many are capable of experiencing same illusions as people (Benhar, Samuel 1982; Kelley, Kelley 2013; Logothetis 1998; Tudusciuc, Nieder 2010), while also experiencing our world in a very different way (Lazareva *et al.* 2012). Historically, we have been greatly underestimating consciousness of animals (Low *et al.* 2012), and it is likely that now we are doing it to intelligent machines.

What it feels like to be a particular type of agent in a given situation depends on the hardware/software/state of the agent and stimulation being provided by the environment. As the qualia represent the bedrock of consciousness, we can formally define a conscious agent as one capable of experiencing at least some broadly defined illusions. To more formally illustrate this we can represent the agent and its inputs as two shares employed in visual cryptography (Naor, Shamir 1994), depending on the composition of the agent the input may end up producing a diametrically opposite experience (Abboud *et al.* 2010; Yampolskiy *et al.* 2014). Consequently, consciousness is an ability to experience, and we can state two ways in which illusions, and consciousness may interact to produce a conscious agent:

– An agent is real and is experiencing an illusion. This explains qualia and the agent itself is real.

– An agent is real and is having an illusion in which some other agent experiences an illusion. Self-identifying with such an agent creates self-consciousness. A sequence of such episodes corresponds to a stream of consciousness and the illusionary agent itself is not real. You are an illusion experiencing an illusion.

2. TEST FOR DETECTING QUALIA

Illusions provide a tool (Eagleman 2001; Panagiotaropoulos *et al.* 2012), which makes it possible to sneak a peek into the mind of another agent and determine that an agent has in fact experienced an illusion. The approach is similar to non-interactive CAPTCHAs, in which some information is encoded in a CAPTCHA challenge (Ahn *et al.* 2003; D'Souza *et al.* 2012; Korayem *et al.* 2012a, 2012b; Yampolskiy 2012) and it is only by solving the CAPTCHA correctly that the agent is able to obtain information necessary to act intelligently in the world, without having to explicitly self-report its internal state (McDaniel, Yampolskiy 2011, 2012; Yampolskiy 2007; Yampolskiy, Govindaraju 2007). With illusions, it is possible to set up a test in which it is only by experiencing an illusion that the agent is able to enter into a certain internal state, which we can say it experiences. It is not enough to know that something is an illusion. For example, with a classical face/vase illusion (Hasson *et al.* 2001) an agent who was previously not exposed to that challenge, could be asked to report what two interpretations for the image it sees and if the answer matches that of a human experiencing that illusion the agent must also be experiencing the illusion, but perhaps in a different way.

Our proposal represents a variant of a Turing Test (Yampolskiy 2013; Yampolskiy 2012) but with emphasis not on behavior or knowledge but on experiences, feelings and internal states. In related research, Schweizer (Schweizer 2012) has proposed a Total Turing Test for Qualia (Q3T), which is a variant of Turing Test for a robot with sensors and questions concentrated on experiences such as: how do you find that wine? Schneider and Turner have proposed a behavior based AI consciousness test, which looks at whether the synthetic mind has an experience-based understanding of the way it feels to be conscious as demonstrated by an agent «talking» about consciousness related concepts such as afterlife or soul (Schneider, Turner 2017).

What we describe is an empirical test for presence of some subjective experiences. The test is probabilistic but successive different variants of the test can be used to obtain any desired level of confidence. If a collaborating agent fails a particular instance of the test it doesn't mean that the agent doesn't have qualia, but passing an instance of the test should increase our belief that the agent has experiences in proportion to the chance of guessing correct answer for that particular variant of the test. As qualia are agent type (hardware) specific (human, specie, machine, etc.) it would be easiest for us to design a human-compatible qualia test, but in principle, it is possible to test for any type of qualia, even the ones that humans don't experience themselves. Obviously, having some qualia doesn't mean ability to experience them all. While what we propose is a binary detector test for some qualia, it is possible to design specific variants for extracting particular properties of qualia experience such as color, depth, size, etc. The easiest way to demonstrate construction of our test is by converting famous visual illusions into instances of our test questions as seen in Figure 1. Essentially we present our subject with an illusion and ask it a multiple choice question about the illusionary experience, such as: how many black dots do you see? How many curved lines are in the image? Which of the following effects do you observe? It is important to only test subjects with tests they have not experienced before and information about which is not readily available. Ideally a new test question should be prepared every time to prevent the subject from cheating. A variant of the test may ask open ended questions such as: please describe what you see. In that case, a description could be compared to that produced by a conscious agent, but this is less formal and opens the door for subjective interpretation of submitted responses. Ideally, we want to be able to automatically design novel illusions with complex information encoded in them as experiences.

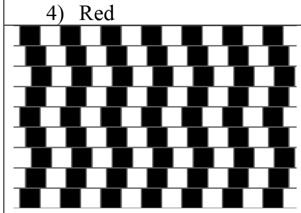
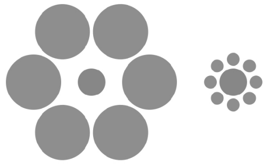
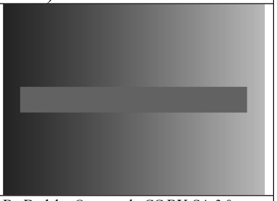
Horizontal lines are: 1) Not in the image 2) Crooked 3) Straight 4) Red	**Orange circles are:** 1) Left one is bigger 2) Right one is bigger 3) They are the same size 4) Not in the image	**Horizontal stripe is:** 1) Solid 2) Spectrum of gray 3) Not in the image 4) Crooked
		
By Fibonacci - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1788689	Public Domain, https://commons.wikimedia.org/w/index.php?curid=828098	By Dodek - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1529278

FIG. 1. Visual Illusions presented as tests.

We anticipate a number of possible objections to the validity of our test and its underlying theory:

– *Qualia experienced by the test subject may not be the same as experienced by the test designer.*

Yes, we are not claiming that they are identical experiences; we are simply showing that an agent had some subjective experiences, which was previously not possible. If sufficiently different, such alternative experiences would not result in passing of the test.

– *The system may simply have knowledge of the human mental model and predict what a human would experience on similar stimulus.*

If a system has an internal human (or some other) model which it simulates on presented stimuli and that generates experiences, it is the same as the whole system having experiences.

– *Agent may correctly guess answers to the test or lie about what it experiences.*

Yes, for a particular test question, but the test can be given as many times as necessary to establish statistical significance.

– *The theory makes no predictions.*

We predict that computers built to emulate the human brain will experience progressively more illusions without being explicitly programmed to do so, in particular the ones typically experienced by people.

Turing addressed a number of relevant objections in his seminar paper on computing machinery (Turing 1950).

3. COMPUTERS CAN EXPERIENCE ILLUSIONS AND SO ARE CONSCIOUS

Majority of scholars studying illusionism are philosophers, but a lot of relevant work comes from psychology (Robinson 2013), cognitive science (Yamins, DiCarlo 2016) and more recently computer science, artificial intelligence, machine learning and more particularly Artificial Neural Network research. It is this interdisciplinary nature of consciousness research which we think is most likely to produce successful and testable theories, such as the theory presented in this paper, to solve the Hard problem.

In the previous section, we have established that consciousness is fundamentally based on an ability to experience, for example illusions. Recent work with artificially intelligent systems suggests that computers also experience illusions and in a similar way to people, providing support for the Principle of Organizational Invariance (Chalmers 1995) aka substrate independence (Bostrom 2003). For example, Zeman *et al.* (2014; 2013) and Garcia-Garibay (García-Garibay, de Lafuente 2015) report on a neural networks capable of experiencing Müller-Lyer illusion and multiple researchers (Bertulis, Bulatov 2001; Corney, Lotto 2007; Inui *et al.* 1990; Ogawa *et al.* 1999) have performed experiments in which computer models were used to study visual illusions, including teaching computers to experience geometric illusions (Chao *et al.* 1993; Ogawa *et al.* 1999), brightness illusions (Robinson *et al.* 2007; Zeman *et al.* 2015) and color constancy illusions (Shibata, Kurizaki 2012). In related research, Nguyen *et al.* found that NN perceive certain random noise images as meaningful with very high confidence (Nguyen *et al.* 2015). Those NN were not explicitly designed to perceive illusions but they do so as a byproduct of the computations they perform. The field of Adversarial Neural Networks is largely about designing illusions for such intelligent systems (Kurakin *et al.* 2016; Szegedy *et al.* 2013) with obvious parallels to inputs known to fool human agents intentionally (Goodfellow *et al.* 2014) or unintentionally (Carlotto 1997). Early work on artificial Neural Networks, likewise provides evidence for experiences similar to near death hallucinations (Thaler 1993; Thaler 1995b) (based on so-called «virtual inputs» or «canonical hallucination» or «neural forgery» (Thaler 1995a), dreaming (Crick, Mitchison 1983; Hopfield *et al.* 1983), and impact from brain damage (Hinton *et al.* 1993; Lecun *et al.* 1990).

Zeman (2015) reviews history of research on perception of illusions by computer models and summarizes the state-of-the-art in such research:

> Historically, artificial models existed that did not contain multiple layers but were still able to demonstrate illusory bias. These models were able to produce output similar to human behaviour when presented with illusory figures, either by emulating the filtering operations of cells (Bertulis, Bulatov 2001; 2005) or by analysing statistics in the environment (Corney, Lotto 2007; Howe, Purves 2002, 2005a, 2005b). However, these models were deterministic, non-hierarchical systems that did not involve any feature learning. It was not until Brown and Friston (2012) that hierarchical systems were first considered as candidates for modelling illusions, even though the authors omitted important details of the model's architecture, such as the number of layers they recruited. […] So to summarise, illusions can manifest in artificial systems that are both hierarchical and

> capable of learning. Whether these networks rely on exposure to the same images that we see during training, or on filtering mechanisms that are based on similar neural operations, they produce a consistent and repeatable illusory bias. In terms of Marr's (1982) levels of description […], it appears that illusions can manifest at the hardware level (Howe, Purves 2005a; 2005b) and at the algorithmic/representational level (Bertulis, Bulatov 2001; 2005; Zeman *et al.* 2013).

«By dissociating our sensory percepts from the physical characteristics of a stimulus, visual illusions provide neuroscientists with a unique opportunity to study the neuronal mechanisms underlying […] sensory experiences» (García-Garibay, de Lafuente 2015). Not surprisingly artificial neural networks just like their natural counterparts are subject to similar analysis. From this, we have to conclude that even today's simple AIs, as they experience specific types of illusions, are rudimentary conscious. General intelligence is what humans have and we are capable of perceiving many different types of complex illusions. As AIs become more adept at experiencing complex and perhaps multisensory illusions they will eventually reach and then surpass our capability in this domain producing multiple parallel streams of superconsciousness (Torrance 2012), even if their architecture or sensors are not inspired by the human brain. Such superintelligent and superconscious systems could justifiably see us as barely intelligent and weakly conscious, and could probably control amount of consciousness they had, within some range. Google deep dream art (Mordvintsev *et al.* 2015) gives us some idea on what it's like to be a modern deep neural network and can be experienced in immersive 3D via the Hallucination Machine (Suzuki *et al.* 2017). Olah *et al.* provide a detailed neuron/layer visual analysis of what is being perceived by an artificial neural network (Olah *et al.* 2017).

3.1. QUALIA COMPUTING

If we can consistently induce qualia in computational agents, it should be possible to use such phenomena to perform computation. If we can encode information in illusions, certain agents can experience them or their combinations to perform computation, including artificially intelligent agents capable of controlling their illusions. Illusions are particularly great to represent superpositions of states (similar to quantum computing), which collapse once a particular view of the illusion is chosen by the experiencing agent (Seckel 2004). You can only experience one interpretation of an illusion at a time, just like in Quantum physics you can only know location or speed of a particle at the same time – well known conjugate pairs (Yampolskiy 2017). Famous examples of logical paradoxes can be seen as useful for super-compressed data storage (Chaitin 1995; Yampolskiy 2013c) and hyper-computation (Potgieter 2006). Qualia may also be useful in explaining decisions produced by deep NN, with the last layer efficiently representing qualia-like states derived from low-level stimuli by lower level neurons. Finally, qualia based visualization and graphics are a very interesting area of investigation, with the human model giving us an example of visual thinking and lucid dreaming.

4. PURPOSE OF CONSCIOUSNESS

While many scientific theories, such as biocentrism (Lanza, Berman 2010) or some interpretations of quantum physics (Goswami 1990; Mould 1998), see consciousness as a focal element of their models, the purpose of being able to experience remains elusive. In fact, even measurement or detection of consciousness remains an open research area (Raoult, Yampolskiy 2015). In this section, we review and elaborate on some explanations for what consciousness does. Many explanations have been suggested, including but certainly not limited to (Blackmore 2016): error monitoring (Crook 1980), an inner eye (Humphrey 1986), saving us from danger (Baars 1997), later error detection (Gray 2004), pramodular response (Morsella 2005) and to seem mysterious (Humphrey 2006).

We can start by considering the evolutionary origins of qualia from the very first, probably accidental, state of matter, that experienced something, all the way to general illusion experiences of modern humans. The argument is that consciousness evolved because accurately representing reality is less important than agents' fitness for survival and agents who saw the world of illusions had higher fitness, as they ignored irrelevant and complicated minutia of the world (Gefter, Hoffman 2016). It seems that processing real world is computationally expensive and simplifying illusions allow improvements in efficiency of decision-making leading to higher survival rates. For example, we can treat feelings as heuristic shortcuts to calculating precise utility. Additionally, as we argue in this paper, experiencing something allows one to obtain knowledge about that experience, which is not available to someone not experiencing the same qualia. Therefore, a conscious agent would be able to perform in ways a philosophical zombie would not be able to act, which is particularly important in the world full of illusions such as ours.

Next, we can look at the value of consciousness in knowledge acquisition and learning. A major obstacle to the successful development of AI systems, has been what is called the Symbol Grounding problem (Harnad 1990). Trying to explain to a computer one symbol in terms of others does not lead to understanding. For example saying that «mother» is a female parent is no different than saying that x = 7y, and y = 18k and so on. This is similar to a person looking up an unfamiliar word in a foreign language dictionary and essentially ending up with circular definitions of unfamiliar terms. We think, that qualia are used (at least in humans) to break out of this vicious cycle and to permit definitions of words/symbols in terms of qualia. In «*How Helen Keller used syntactic semantics to escape from a Chinese Room*», Rappaport (Rapaport 2006) gives a great example of a human attempting to solve the grounding problem and argues that syntactic semantics are sufficient to resolve it. We argue that it is experiencing the feeling of running water on her hands was what permitted Hellen Keller to map sign language sign for water to the relevant qualia and to begin to understand.

Similarly, we see much of the language acquisition process as mapping of novel qualia to words. By extension, this mapping permits us to explain understanding and limits to transfer of tacit knowledge. Illusion disambiguation can play a part in what gives us an illusion of free will and the stream of consciousness may be nothing more than sequential illusion processing. Finally, it would not be surprising if some implicit real-world inputs produced experience of qualia behind some observed precognition results (Mossbridge *et al.* 2012). In the future, we suspect a major application of consciousness will be in the field of Qualia Computing as described in the so-named section of this paper.

4.1. QUALIA ENGINEERING

While a grand purpose of life remains elusive and is unlikely to be discovered, it is easy to see that many people attempt to live their lives in a way, which allows them to maximally explore and experience novel stimuli: foods, smells, etc. Experiencing new qualia by transferring our consciousness between different substrates, what Loosemore refers to as Qualia Surfing (Loosemore 2014), may represent the next level in novelty seeking. As our understanding and ability to detect and elicit particular qualia in specific agents improves, qualia engineering will become an important component of the entertainment industry. Research in other fields such as: intellectology (Yampolskiy 2015b), (and in particular artimetrics, Yampolskiy *et al.* 2012; Yampolskiy, Gavrilova 2012), and designometry (Yampolskiy 2016a), consciousness (Yampolskiy 2018) and artificial intelligence (Yampolskiy, Fox 2012) will also be impacted.

People designing optical illusions, movie directors and book authors are some of the people in the business of making us experience, but they do so as an art form. Qualia engineers and qualia designers will attempt to formally and scientifically answer such questions as: How to detect and measure qualia? What is the simplest possible qualia? How to build complex qualia from simple ones? What makes some qualia more pleasant? Can minds be constructed with maximally pleasing qualia in a systematic and automated way (Yampolskiy 2015b)? Can this lead to abolition of suffering (Hughes 2011)? Do limits exist to complexity of qualia, or can the whole universe be treated as single input? Can we create new feelings and emotions? How would integration of novel sensors expand our qualia repertoire? What qualia are available to other agents but not to humans? Can qualia be «translated» to other mediums? What types of verifiers and observers experience particular types of qualia? How to generate novel qualia in an algorithmic/systematic way? Is it ethical to create unpleasant qualia? Can agents learn to swap qualia between different stimuli (pleasure for pain)? How to optimally represent, store and communicate qualia, including across different substrates (Bostrom 2003)? How to design an agent, which experiences particular qualia on the given input? How much influence does an agent have over its own illusions? How much plasticity does the human brain have for switching stimuli streams and learning to experience data from new sensors? How similar are qualia among similarly designed but not identical agents? What, if any, is the connection between meditation and qualia? Can computers mediate? How do random inputs such as café chatter (Mehta *et al.* 2012) stimulate production of novel qualia? How can qualia be classified into different types, for example feelings? Which computations produce particular qualia?

5. CONSCIOUSNESS AND ARTIFICIAL INTELLIGENCE

Traditionally, AI researchers ignored consciousness as non-scientific and concentrated on making their machines capable and beneficial. One famous exception is Hofstadter who observed and analyzed deep connections between illusions and artificial intelligence (Hofstadter 1979). If an option to make conscious machines presents itself to AI researchers, it would raise a number of important questions, which should be addressed early on. It seems that making machines conscious may make them more relatable and human like and so produce better consumer products, domestic and sex robots and more genuine conversation partners. Of course, a system simply simulating such behaviors without actually experiencing anything could be just as good. If we define physical pain as an unpleasant sensory illusion and emotional pain as an illusion of an unpleasant feeling, pain and pleasure become accessible controls to the experimenter. Ability to provide reward and punishment for software agents capable of experiencing pleasure and pain may assist in the training of such agents (Majot, Yampolskiy 2014).

Potential impact from making AI conscious includes change in the status of AI from mere useful software to a sentient agent with corresponding rights and ethical treatment standards. This is likely to lead to civil rights for AI and disenfranchisement of human voters (Yampolskiy 2013a; 2013b). In general, ethics of designing sentient beings are not well established and it is cruel to create sentient agents for certain uses, such as menial jobs, servitude or designed obsolescence. It is an experiment which would be unlikely to be approved by any research ethics board (Braverman 2017). Such agents may be subject to abuse as they would be capable of experiencing pain and torture, potentially increasing the overall amount of suffering in the universe (Metzinger 2017). If in the process of modeling or simulating conscious beings, experiment negatively affects modeled entities this can be seen as mind crime (Bostrom 2014).

With regards to AI safety (Babcock *et al.* 2016; 2017; Pistono, Yampolskiy 2016; Yampolskiy 2015a; 2016b; Yampolskiy, Spellchecker 2016), since it would be possible for agents to experience pain and pleasure it will open a number of new pathways for dangerous behavior. Consciousness may make AIs more volatile or unpredictable impacting overall safety and stability of such systems (Schneider, Turner, 2017). Possibility of ransomware with conscious artificial hostages comes to mind as well as blackmail and threats against AI system. Better understanding of consciousness by AI itself may also allow superintelligent machines to create new types of attacks on people. Certain illusions can be seen as an equivalent of adversarial inputs for human agents, see Figure 2. Subliminal stimuli (Greenwald *et al.* 1995) which confuse people are well known and some stimuli are even capable of inducing harmful internal states such as epileptic seizures (Harding, Jeavons 1994; Walter *et al.* 1946) or incapacitation (Altmann 2001). With latest research showing, that even a single pixel modification is sufficient to fool neural networks (Su *et al.* 2017), the full scope of the attack surface against human agents remains an unknown unknown.

Manual attempts to attack a human cognitive model are well known (Bandler *et al.* 1982; Barber 1969; Vokey, Read 1985). Future research combining evolutionary algorithms or adversarial neural networks with direct feedback from detailed scans of human brains is likely to produce some novel examples of adversarial human inputs, leading to new types of informational hazards (Bostrom 2011). Taken to the extreme, whole adversarial worlds may be created to confuse us (Bostrom 2003). Nature provides many examples of adversarial inputs in plants and animals, known as mimicry (Wickler 1968). Human adversarial inputs designed by superintelligent machines would represent a new type of AI risk, which has not been previously analyzed and with no natural or synthetic safety mechanisms available to defend us against such an attack.

One very dangerous outcome from integration of consciousness into AI is a possibility that a superintelligent system will become a negative utilitarian and an anti-natalist (Metzinger 2017) and in an attempt to rid the world of suffering will not only kill all life forms, but will also destroy all AIs and will finally self-destruct as it is itself conscious and so subject to the same analysis and conclusions. This would result in a universe free of suffering but also free of any consciousness. Consequently, it is important to establish guidelines and review boards (Yampolskiy, Fox 2013) for any research which is geared at producing conscious agents (Yampolskiy 2017). AI itself should be designed to be corrigible (Soares *et al.* 2015) and to report any emergent un-programmed capabilities, such as qualia, to the designers.

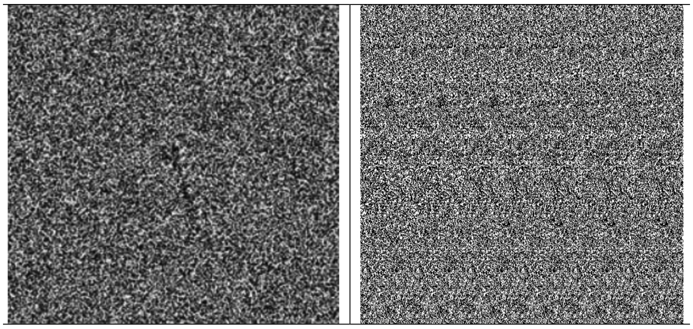


FIG. 2. *Left – Cheetah in the noise is seen by some Deep Neural Networks (based on Nguyen et al. 2015); Right – Spaceship in the stereogram is seen by some people.*

6. CONCLUSIONS AND CONJECTURES

In this paper, we described a reductionist theory for appearance of qualia in agents based on a fully materialistic explanation for subjective states of mind, an attempt at a solution to the Hard Problem of consciousness. We defined a test for detecting experiences and showed how computers can be made conscious in terms of having qualia. Finally, we looked at implications of being able to detect and generate qualia in artificial intelligence. Should our test indicate presence of complex qualia in software or animals certain protections and rights would be appropriate to grant to such agents. Experimental results, we surveyed in this paper, have been predicted by others as evidence of consciousness in machines, for example Dehaene *et al.* state: «We contend that a machine endowed with [global information availability and self-monitoring] […] may even experience the same perceptual illusions as humans» (Dehaene *et al.* 2017).

Subjective experiences called qualia are a side effect of computing, unintentionally produced while information is being processed, similar to generation of heat (Landauer 1961), noise (Genkin *et al.* 2014), or electromagnetic radiation (De Mulder *et al.* 2006) and is just as unintentional. Others have expressed similar intuitions: «The cognitive algorithms we use are the way the world feels.» (Yudkowsky 2015, p. 889) or «consciousness is the way information feels when being processed.» (Hut *et al.* 2006) or «empirical evidence is compatible with the possibility that consciousness arises from nothing more than specific computations» (Dehaene *et al.* 2017). Qualia arise as a result of processing of stimuli caused by agglomeration of properties, unique peculiarities (Schwarting *et al.* 2015) and errors in agent's architecture, software, memories, learned algorithms, sensors, inputs, environment and other factors comprising extended cognition (Rupert 2004) of an agent (Clark, Chalmers 1998). In fact, Zeman (2015) points out the difficulty of telling if a given system experiences an error or an illusion. If every computation produces side effect of qualia, computational functionalism (Putnam 1980) trivially reduces to panpsychism (Chalmers 1996).

As qualia are fully dependent on a makeup of a particular agent it is not surprising that they capture what it is like to be that agent. Agents, which share certain similarities in their makeup (like most people), may share certain subsets of qualia, but different agents will experience different qualia on the same inputs. An illusion is a discrepancy between agent's awareness and some stimulus (Reynolds 1988). In contrast, consciousness is an ability to experience a sustained self-referential multimodal illusion based on an ability to perceive qualia. Every experience is an illusion, what we call optical illusions are meta illusions, there are also meta-meta-illusions and self-referential illusions. It is an illusion of «I» or self which produces self-awareness, with «I» as an implied agent experiencing all the illusions, an illusion of an illusion navigator.

It is interesting to view the process of learning in the context of this paper, with illusions as a primary pattern of interest for all agents. We can say that babies and other untrained neural networks are learning to experience illusions, particularly in the context of their trainers' culture/common sense (Segall *et al.* 1963). Consequently, a successful agent will learn to map certain inputs to certain illusions while sharing that mapping with other similarly constructed observers. We can say that the common space of illusions/culture as seen by such agents becomes their «real world» or meme (Dawkins 1976) sphere. Some supporting evidence for this conclusion comes from observing that amount of sleep in children is proportionate to the average amount of learning they perform for that age group. Younger babies need the most sleep, perhaps because they can learn quicker by practicing to experience in the safe world of dreams (a type of illusion) a skill they then transfer to the real world. Failure to learn to perceive illusions and experience qualia may result in a number of mental disorders.

There seems to be a fundamental connection between intelligence, consciousness and liveliness beyond the fact that all three are notoriously difficult to define. We believe that ability to experience is directly proportionate to one's intelligence and that such intelligent and conscious agents are necessarily alive to the same degree. As all three come in degrees, it is likely that they have gradually evolved together. Modern narrow AIs are very low in general intelligence and so are also very low in their ability to experience or their perceived liveness. Higher primates have significant (but not complete) general intelligence and so can experience complex stimuli and are very much alive. Future machines will be superintelligent, superconscious and by extension alive!

Fundamental «particles» from which our personal world is constructed are illusions, which we experience and in the process create the universe, as we know it. Experiencing a pattern which is not really there (let's call such an illusory element «illusination«), like appearing white spaces in an illusion (Ninio, Stevens 2000), is just like experiencing self-awareness; where is it stored? Since each conscious agent perceives a unique personal universe, their agglomeration gives rise to the multiverse. We may be living in a simulation, but from our point of view we are not living in a virtual reality (Chalmers 2016), we are living in an illusion of reality, and maybe we can learn to decide which reality to create. The «Reality» provides us with an infinite set of inputs from which every conceivable universe can be experienced and in that sense, every universe exists. We can conclude that the universe is in the mind of the agent experiencing it - the ultimate qualia, even if we are just brains in a vat, to us an experience is worth a 1000 pictures. It is not a delusion that we are just experiencers of illusions. Brain is an illusion experiencing machine not a pattern recognition machine. As we age, our wetware changes and so we become different agents and experience different illusion, our identity changes but in a continuous manner. To paraphrase Descartes: I experience, therefore I am conscious!

Roman V. Yampolskiy
Computer Engineering and Computer Science
Speed School of Engineering
University of Louisville
roman.yampolskiy@louisville.edu

ACKNOWLEDGEMENTS

The author is grateful to Elon Musk and the Future of Life Institute and to Jaan Tallinn and Effective Altruism Ventures for partially funding his work on AI Safety. The author is thankful to Yana Feygin for proofreading a draft of this paper and to Ian Goodfellow for helpful recommendations of relevant literature.

ENDNOTES

[1] Kolmogorov complexity is also not computable, but very useful.

REFERENCES

Abboud G., Marean J., Yampolskiy R.V. (2010), *Steganography and Visual Cryptography in Computer Forensics,* Paper presented at the Systematic Approaches to Digital Forensic Engineering (SADFE), 2010 Fifth IEEE International Workshop on.

Ahn L. v., Blum M., Hopper N., Langford J. (2003), *CAPTCHA: Using Hard AI Problems for Security,* Paper presented at the Eurocrypt.

Altmann J. (2001), *Acoustic Weapons. A Prospective Assessment*, in «Science & Global Security», 9(3), 165-234.

Baars B.J. (1997), *In the Theatre of Consciousness. Global Workspace Theory, a Rigorous Scientific Theory of Consciousness*, in «Journal of Consciousness Studies», 4(4), 292-309.

Babcock J., Kramár J., Yampolskiy R. (2016), *The AGI Containment Problem*, Paper presented at the International Conference on Artificial General Intelligence.

Babcock J., Kramar J., Yampolskiy R.V. (2017), *Guidelines for Artificial Intelligence Containment*, arXiv preprint arXiv:1707.08476.

Bach-y-Rita P., Kercel S.W. (2003), *Sensory Substitution and the Human-Machine Interface*, in «Trends in cognitive sciences», 7(12), 541-546.

Balog K. (2016), *Illusionism's Discontent*, in «Journal of Consciousness Studies», 23(11-12), 40-51.

Bancaud J., Brunet-Bourgin F., Chauvel P., Halgren E., Bancaud T. (1994), *Anatomical Origin of Déjà Vu and Vivid «Memories» in Human Temporal Lobe Epilepsy*, in «Brain», 117(1), 71-90.

Bandler R., Grinder J., Andreas S. (1982), *Neuro-Linguistic Programming™ and the Transformation of Meaning*, Real People, Moab.

Barber T. X. (1969), *Hypnosis: A scientific approach*, Oxford, Van Nostrand Reinhold.

Barrett D. (1992), *Just how Lucid are Lucid Dreams?*, in «Dreaming», 2(4), 221.

Becker H.S. (1967), *History, Culture and Subjective Experience: An Exploration of the Social Bases of Drug-Induced Experiences*, in «Journal of Health and Social Behavior», 163-176.

Becker S., Hinton G.E. (1992), *Self-Organizing Neural Network that Discovers Surfaces in Random-Dot Stereograms*, in «Nature», 355(6356), 161-163.

Benhar E., Samuel D. (1982), *Visual Illusions in the Baboon (Papio Anubis)*, in «Learning & Behavior», 10(1), 115-118.

Bentall R.P. (1990), *The Illusion of Reality: A Review and Integration of Psychological Research on Hallucinations*, in «Psychological Bulletin», 107(1), 82.

Bertamini M. (2017), *Programming Visual Illusions for Everyone*, Berlin, Springer.

Bertulis A., Bulatov A. (2001), *Distortions of Length Perception in Human Vision*, in «Biomedicine», 1(1), 3-23.

Bertulis A., Bulatov A. (2005), *Distortions in Length Perception: Visual Field Anisotropy and Geometrical Illusions*, in «Neuroscience and Behavioral Physiology», 35(4), 423-434.

Blackmore S. (2016), *Delusions of Consciousness*, in «Journal of Consciousness Studies», 23(11-12), 52-64.

Block N. (1995), *On a Confusion about a Function of Consciousness*, in «Behavioral and Brain Sciences», 18(2), 227-247.

Bostrom N. (2003), *Are we Living in a Computer Simulation?*, in «The Philosophical Quarterly», 53(211), 243-255.

Bostrom N. (2011), *Information Hazards: A Typology of Potential Harms from Knowledge*, in «Review of Contemporary Philosophy», 10, 44.

Bostrom N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford, Oxford University Press.

Braverman I. (2017), *Gene Drives, Nature, Governance: An Ethnographic Perspective*, Paper presented at the University at Buffalo School of Law Legal Studies Research Paper No. 2017-006, Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3032607.

Brown H., Friston K.J. (2012), *Free-Energy and Illusions: The Cornsweet Effect*, in «Frontiers in Psychology», 3.

Burn C.C. (2008), *What Is it Like to Be a Rat? Rat Sensory Perception and its Implications for Experimental Design and Rat Welfare*, in «Applied Animal Behaviour Science», 112(1), 1-32.

Carlen P., Wall, P., Nadvorna, H., Steinbach, T. (1978), *Phantom Limbs and Related Phenomena in Recent Traumatic Amputations*, in «Neurology», 28(3), 211-211.

Carlotto M. J. (1997), *The Martian Enigmas: A Closer Look: The Face, Pyramids and Other Unusual Objects on Mars*, Berekeley (CA), North Atlantic Books.

Chaitin G.J. (1995), *The Berry Paradox*, in «Complexity», 1(1), 26-30.

Chalmers D.J. (1993), *Self-Ascription Without Qualia: A Case Study*, in «Behavioral and Brain Sciences», 16(1), 35-36.

Chalmers D.J. (1995), *Facing up to the Problem of Consciousness*, in «Journal of Consciousness Studies», 2(3), 200-219.

Chalmers D.J. (1996), *Does a Rock Implement Every Finite-State Automaton?*, in «Synthese», 108(3), 309-333.

Chalmers D. J. (2016), *The virtual and the real*, in «Disputatio», 9(46), 309-352.

Changizi M.A., Hsieh A., Nijhawan R., Kanai R., Shimojo S. (2008), *Perceiving the Present and a Systematization of Illusions*, in «Cognitive science», 32(3), 459-503.

Chao J., Kishigami T., Minowa K., Tsujii S. (1993), *Artificial Neural Networks which Can See Geometric Illusions in Human Vision*, Paper presented at the Neural Networks, 1993. IJCNN'93-Nagoya. Proceedings of 1993 International Joint Conference on.

Clark A., Chalmers D. (1998), *The Extended Mind*, in «Analysis», 58(1), 7-19.

Coren S., Girgus J. S. (1978), *Seeing is deceiving: The psychology of visual illusions*, Mahwah (NJ), Lawrence Erlbaum..

Corney D., Lotto R.B. (2007), *What are Lightness Illusions and Why Do We See Them?*, in «PLoS computational Biology», 3(9), e180.

Coslett H.B., Saffran E. (1991), *Simultanagnosia: To See but not Two See*, in «Brain», 114(4), 1523-1545.

Crick, F., Mitchison G. (1983), *The Function of Dream Sleep*, in «Nature», 304(5922), 111-114.

Crook J.H. (1980), *The Evolution of Human Consciousness*, Oxford, Clarendon Press.

Cytowic R.E. (2002), *Synesthesia: A Union of the Senses*, Cambridge, MIT Press.

D'Souza D., Polina P.C., Yampolskiy R.V. (2012), *Avatar CAPTCHA: Telling Computers and Humans Apart Via Face Classification*, Paper presented at the Electro/Information Technology (EIT), 2012 IEEE International Conference on.

Damasio A.R., Damasio H., Van Hoesen G.W. (1982), *Prosopagnosia Anatomic Basis and Behavioral Mechanisms*, in «Neurology», 32(4), 331-331.

Dawkins R. (1976), *The Selfish Gene,* New York, Oxford University Press.

De Mulder E., Ors S., Preneel B., Verbauwhede I. (2006), *Differential Electromagnetic Attack on an FPGA Implementation of Elliptic Curve Cryptosystems*, Paper presented at the Automation Congress, 2006, WAC'06. World.

Dehaene S., Lau H., Kouider S. (2017), *What is Consciousness, and Could Machines Have It?*, in «Science», 358(6362), 486-492.

Dennett D.C. (1981), *Brainstorms: Philosophical Essays on Mind and Psychology*, Cambridge, MIT Press.

Dennett D.C. (2017), *From Bacteria to Bach and Back: The Evolution of Minds*, New York, W.W. Norton & Co.

Deręgowski J.B. (2015), *Illusions Within an Illusion*, in «Perception», 44(12), 1416-1421.

Deutsch D. (1974), *An Auditory Illusion*, in «The Journal of the Acoustical Society of America», 55(S1), S18-S19.

Dima D., Roiser J.P., Dietrich D.E., Bonnemann C., Lanfermann H., Emrich H.M., Dillo W. (2009), *Understanding why Patients with Schizophrenia Do not Perceive the Hollow-Mask Illusion Using Dynamic Causal Modelling*, in «Neuroimage», 46(4), 1180-1186.

Eagleman D.M. (2001), *Visual Illusions and Neurobiology*, in «Nature Reviews Neuroscience», 2(12), 920-926.

Eagleman D.M. (2008), *Human Time Perception and Its Illusions*, in «Current opinion in neurobiology», 18(2), 131-136.

Ehrsson H.H. (2007), *The Experimental Induction of Out-of-Body Experiences*, in «Science», 317(5841), 1048-1048.

Escher M.C. (2000), *MC Escher: The Graphic Work*, Berlin, Taschen.

Fenwick P. (1996), *The Neurophysiology of Religious Experiences*, London, Routledge.

Fineman M. (2012), *The Nature of Visual Illusion*, Chelmsford, Courier Corporation.

Frankish K. (2016), *Illusionism as a Theory of Consciousness*, in «Journal of Consciousness Studies», 23(11-12), 11-39.

Fyfe S., Williams C., Mason O.J., Pickup G.J. (2008), *Apophenia, Theory of Mind and Schizotypy: Perceiving Meaning and Intentionality in Randomness*, in «Cortex», 44(10), 1316-1325.

García-Garibay O.B., de Lafuente V. (2015), *The Müller-Lyer Illusion as Seen by an Artificial Neural Network*, in «Frontiers in Computational Neuroscience», 9.

Garety P.A., Hemsley D.R. (1997), *Delusions: Investigations into the Psychology of Delusional Reasoning*, New York, Psychology Press.

Gefter A., Hoffman D.D. (2016), *The Evolutionary Argument Against Reality*, Quanta Magazine.

Genkin D., Shamir A., Tromer E. (2014), *RSA Key Extraction Via Low-Bandwidth Acoustic Cryptanalysis*, Paper presented at the International Cryptology Conference.

Gigerenzer G. (1991), *How to Make Cognitive Illusions Disappear: Beyond «Heuristics and Biases»*, in «European Review of Social Psychology», 2(1), 83-115.

Gillespie A. (2006), *Descartes' Demon: A Dialogical Analysis of Meditations on First Philosophy*, in «Theory & Psychology», 16(6), 761-781.

Gold R. (1993), *This Is not a Pipe*, Communications of the ACM, 36(7), 72.

Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Bengio Y. (2014), *Generative Adversarial Nets*, Paper presented at the Advances in Neural Information Processing Systems.

Goswami A. (1990), *Consciousness in Quantum Physics and the Mind-Body Problem*, in «The Journal of Mind and Behavior», 75-96.

Gray C.H. (2000), *Cyborg Citizen: Politics in the Posthuman Age*, London, Routledge.

Gray J.A. (2004), *Consciousness: Creeping up on the Hard Problem*, Oxford, Oxford University Press.

Greenleaf A., Kurylev Y., Lassas M., Uhlmann G. (2011), *Schrödinger's Hat: Electromagnetic, Acoustic and Quantum Amplifiers Via Transformation Optics*, arXiv preprint arXiv:1107.4685.

Greenwald A.G., Klinger M.R., Schuh E.S. (1995), *Activation by Marginally Perceptible («subliminal») Stimuli: Dissociation of Unconscious from Conscious Cognition*, in «Journal of Experimental Psychology: General», 124(1), 22.

Gregory R.L. (1997), *Knowledge in Perception and Illusion*, in «Philosophical Transactions of the Royal Society of London B: Biological Sciences», 352(1358), 1121-1127.

Grelling K. (1936), *The Logical Paradoxes*, in «Mind», 45(180), 481-486.

Grube G.M.A. (1974), *Plato's republic*, Indianapolis, Hackett Publishing.

Happé F.G. (1996), *Studying Weak Central Coherence at Low Levels: Children with Autism Do not Succumb to Visual Illusions, A Research Note*, in «Journal of Child Psychology and Psychiatry», 37(7), 873-877.

Harding G.F., Jeavons P.M. (1994), *Photosensitive Epilepsy*, Cambridge, Cambridge University Press.

Harlow H.F., Stagner R. (1932), *Psychology of Feelings and Emotions: I. Theory of Feelings*, in «Psychological Review», 39(6), 570.

Harnad S. (1990), *The Symbol Grounding Problem*, in «Physica D: Nonlinear Phenomena», 42(1-3), 335-346.

Hasson U., Hendler T., Bashat D.B., Malach R. (2001), *Vase or Face? A Neural Correlate of Shape-Selective Grouping Processes in the Human Brain*, in «Journal of Cognitive Neuroscience», 13(6), 744-753.

Herz R.S., von Clef J. (2001), *The Influence of Verbal Labeling on the Perception of Odors: Evidence for Olfactory Illusions?*, in «Perception», 30(3), 381-391.

Hinton G.E., Plaut D.C., Shallice T. (1993), *Simulating Brain Damage*, in «Scientific American», 269(4), 76-82.

Hofstadter D.R. (1979), *Gödel, Escher, Bach: An Eternal Golden Braid*, New York, Basic Books.

Hood B. (2012), *The Self Illusion: How the Social Brain Creates Identity*, Oxford, Oxford University Press.

Hopfield J.J., Feinstein D.I., Palmer R.G. (1983), *«Unlearning» Has a Stabilizing Effect in Collective Memories*, in «Nature», 304(5922), 158-159.

Howe C.Q., Purves D. (2002), *Range Image Statistics Can Explain the Anomalous Perception of Length*, in «Proceedings of the National Academy of Sciences», 99(20), 13184-13188.

Howe C.Q., Purves D. (2005a), *The Müller-Lyer Illusion Explained by the Statistics of Image-Source Relationships*, in «Proceedings of the National Academy of Sciences of the United States of America», 102(4), 1234-1239.

Howe C.Q., Purves D. (2005b), *Perceiving Geometry: Geometrical Illusions Explained by Natural Scene Statistics*, Berlin, Springer.

Hughes J. (2011), *After Happiness, Cyborg Virtue*, in «Free Inquiry», 32(1), 1-7.

Humphrey N. (1986), *The Inner Eye*, Oxford, Oxford University Press.

Humphrey N. (2006), *Seeing Red: A Study in Consciousness*, Cambridge, Harvard University Press.

Hut P., Alford M., Tegmark M. (2006), *On Math, Matter and Mind*, in «Foundations of Physics», 36(6), 765-794.

Inui T., Hongo S., Kawato M. (1990), *A Computational Model of Brightness Illusion and Its Implementation*, in «Perception», 19, 401.

Izard C.E. (1991), *The Psychology of Emotions*, Berlin, Springer.

Jackson F. (1986), *What Mary Didn't Know*, in «The Journal of Philosophy», 83(5), 291-295.

Jürgens U.M., Nikolić D. (2014), *Synaesthesia as an Ideasthesia – Cognitive Implications*, in «Synaesthesia and Children – Learning and Creativity».

Kahneman D., Tversky A. (1996), *On the Reality of Cognitive Illusions*, in «Psychological Review», 103(3), 582-591.

Keane B.P., Silverstein S.M., Wang Y., Papathomas T.V. (2013), *Reduced Depth Inversion Illusions in Schizophrenia Are State-Specific and Occur for Multiple Object Types and Viewing Conditions*, «Journal of Abnormal Psychology», 122(2), 506.

Kelley L.A., Endler J.A. (2012), *Illusions Promote Mating Success in Great Bowerbirds*, in «Science», 335(6066), 335-338.

Kelley L.A., Kelley J.L. (2013), *Animal Visual Illusion and Confusion: The Importance of a Perceptual Perspective*, in «Behavioral Ecology», 25(3), 450-463.

Kendrick M. (2009), *Tasting the Light: Device Lets the Blind «See» with Their Tongues*, in «Scientific American», 13.

Kluft R.P. (1996), *Dissociative Identity Disorder*, in *Handbook of Dissociation*, Berlin, Springer, 337-366.

Koffka K. (2013), *Principles of Gestalt Psychology*, New York, Routledge.

Korayem M., Mohamed A.A., Crandall D., Yampolskiy R.V. (2012a), *Learning Visual Features for the Avatar Captcha Recognition Challenge*, Paper presented at the Machine Learning and Applications (ICMLA), 2012 11th International Conference on.

Korayem, M., Mohamed A.A., Crandall D., Yampolskiy R.V. (2012b), *Solving Avatar Captchas Automatically*, in *Advanced Machine Learning Technologies and Applications*, Berlin, Springer, 102-110.

Kurakin A., Goodfellow I., Bengio S. (2016), *Adversarial Examples in the Physical World,* arXiv preprint arXiv:1607.02533.

Landauer R. (1961), *Irreversibility and Heat Generation in the Computing Process*, in «IBM Journal of Research and Development», 5(3), 183-191.

Lanza R., Berman B. (2010), *Biocentrism: How Life and Consciousness Are the Keys to Understanding the True Nature of the Universe*, BenBella Books.

Laureys S., Boly M. (2007), *What Is It Like to Be Vegetative or Minimally Conscious?*, in «Current Opinion in Neurology», 20(6), 609-613.

Lazareva O.F., Shimizu T., Wasserman E.A. (2012), *How Animals See the World: Comparative Behavior, Biology, and Evolution of Vision*, Oxford, Oxford University Press.

Lecun Y., Denker J.S., Solla S.A. (1990), *Optimal Brain Damage*, in D. Touretzky (ed.), *Advances in Neural Information Processing Systems (NIPS 1989)*, Denver, Morgan Kaufmann.

Liebe C.C. (1993), *Pattern Recognition of Star Constellations for Spacecraft Applications*, in «IEEE Aerospace and Electronic Systems Magazine», 8(1), 31-39.

Liu J., Li J., Feng L., Li L., Tian J., Lee K. (2014), *Seeing Jesus in Toast: Neural and Behavioral Correlates of Face Pareidolia*, in «Cortex», 53, 60-77.

Logothetis N.K. (1998), *Single Units and Conscious Vision*, in «Philosophical Transactions of the Royal Society of London B: Biological Sciences», 353(1377), 1801-1818.

Loosemore R. (2014), *Qualia Surfing*, in D. Broderick, R. Blackford (eds.), *Intelligence Unbound: Future of Uploaded and Machine Minds*, Hoboken (NJ), Wiley-Blackwell, 231-239.

Lord E. (1950), *Experimentally Induced Variations in Rorschach Performance*, in «Psychological Monographs: General and Applied», 64(10), i.

Low P., Panksepp J., Reiss D., Edelman D., Van Swinderen B., Koch C. (2012), *The Cambridge declaration on consciousness*, Paper presented at the Francis Crick Memorial Conference, Cambridge, England.

Luckiesh M. (1922), *Visual Illusions: Their Causes, Characteristics and Applications*, New York, D. Van Nostrand.

Majot A.M., Yampolskiy R.V. (2014), *AI Safety Engineering Through Introduction of Self-Reference into Felicific Calculus Via Artificial Pain and Pleasure*, Paper presented at the IEEE International Symposium on Ethics in Science, Technology and Engineering, Chicago, IL.

Marr D. (1982), *Vision a Computational Investigation into the Human Representation and Processing of Visual Information*, WH San Francisco, Freeman & Co., 1(2).

McDaniel R., Yampolskiy R.V. (2011), *Embedded Non-Interactive CAPTCHA for Fischer Random Chess*, Paper presented at the 16th International Conference on Computer Games (CGAMES), Louisville, KY.

McDaniel R., Yampolskiy R.V. (2012), *Development of Embedded CAPTCHA Elements for Bot Prevention in Fischer Random Chess*, in «International Journal of Computer Games Technology», 2012, 2.

Mehta R., Zhu R., Cheema A. (2012), *Is Noise Always Bad? Exploring the Effects of Ambient Noise on Creative Cognition*, in «Journal of Consumer Research», 39(4), 784-799.

Mennell S. (1996), *All Manners of Food: Eating and Taste in England and France from the Middle Ages to the Present*, Chicago, University of Illinois Press.

Metzinger T. (2017), *Benevolent Artificial Anti-Natalism (BAAN)*, Paper presented at the EDGE, Available at: https://www.edge.org/conversation/thomas_metzinger-benevolent-artificial-anti-natalism-baan.

Misra B., Sudarshan E.G. (1977), *The Zeno's Paradox in Quantum Theory*, in «Journal of Mathematical Physics», 18(4), 756-763.

Mordvintsev A., Olah C., Tyka M. (2015), *Inceptionism: Going Deeper into Neural Networks*, Google Research Blog. Retrieved June, 20, 14.

Mormann F., Koch C. (2007), *Neural Correlates of Consciousness*, in «Scholarpedia», 2(12), 1740.

Morsella E. (2005), *The Function of Phenomenal States: Supramodular Interaction Theory*, in «Psychological Review», 112(4), 1000.

Mossbridge J., Tressoldi P., Utts J. (2012), *Predictive Physiological Anticipation Preceding Seemingly Unpredictable Stimuli: A Meta-Analysis*, in «Frontiers in Psychology», 3.

Mould R.A. (1998), *Consciousness and Quantum Mechanics*, in «Foundations of Physics», 28(11), 1703-1718.

Muehlhauser L. (2017), *Report on Consciousness and Moral Patienthood,* Paper presented at the Open Philanthropy Project, Available at: https://www.openphilanthropy.org/2017-report-consciousness-and-moral-patienthood.

Nagel T. (1974), *What Is It Like to Be a Bat?*, in «The Philosophical Review», 83(4), 435-450.

Nakatani M., Howe R.D., Tachi S. (2006), *The Fishbone Tactile Illusion*, Paper presented at the Proceedings of Eurohaptics.

Naor M., Shamir A. (1994), *Visual Cryptography*, Paper presented at the Workshop on the Theory and Application of of Cryptographic Techniques.

Nguyen A., Yosinski J., Clune J. (2015), *Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images*, Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Ninio J., Stevens K.A. (2000), *Variations on the Hermann Grid: An Extinction Illusion*, in «Perception», 29(10), 1209-1217.

Noë A. (2002), *Is the Visual World a Grand Illusion?*, in «Journal of Consciousness Studies», 9(5-6), 1-12.

O'Regan J.K. (2011), *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*, Oxford, Oxford University Press.

Ogawa T., Minohara T., Kanada H., Kosugi Y. (1999), *A Neural Network Model for Realizing Geometric Illusions Based on Acute-Angled Expansion*, Paper presented at the Neural Information Processing, 1999, Proceedings. ICONIP'99. 6th International Conference on.

Ogawa T., Minohara T., Kanada H., Kosugi Y. (1999), *Realization of Geometric Illusions Using Artificial Visual Model Based on Acute-Angled Expansion Among Crossing Lines*, Paper presented at the Neural Networks, 1999. IJCNN'99. International Joint Conference on.

Olah C., Mordvintsev A., Schubert L. (2017), *Feature Visualization*, in «Distill», 2(11), e7.

Özkural E. (2012), *What Is It Like to Be a Brain Simulation?*, Paper presented at the International Conference on Artificial General Intelligence.

Panagiotaropoulos T.I., Deco G., Kapoor V., Logothetis N.K. (2012), *Neuronal Discharges and Gamma Oscillations Explicitly Reflect Visual Consciousness in the Lateral Prefrontal Cortex*, in «Neuron», 74(5), 924-935.

Penrose L.S., Penrose R. (1958), *Impossible Objects: A Special Type of Visual Illusion*, in «British Journal of Psychology», 49(1), 31-33.

Pistono F., Yampolskiy R.V. (2016), *Unethical Research: How to Create a Malevolent Artificial Intelligence*, arXiv preprint arXiv:1605.02817.

Post R.H. (1962), *Population Differences in Red and Green Color Vision Deficiency: A Review, and a Query on Selection Relaxation*, in «Eugenics Quarterly», 9(3), 131-146.

Potgieter P.H. (2006), *Zeno Machines and Hypercomputation*, in «Theoretical Computer Science», 358(1), 23-33.

Preuss T.M. (2004), *What Is It Like to Be a Human*, in «The Cognitive Neurosciences», 3, 5-22.

Putnam H. (1980), *The Nature of Mental States*, in «Readings in Philosophy of Psychology», 1, 223-231.

Raoult A., Yampolskiy R. (2015), *Reviewing Tests for Machine Consciousness*, Available at: https://www.researchgate.net/publication/284859013_DRAFT_Reviewing_Tests_for_Machine_Consciousness.

Rapaport, W.J. (2006), *How Helen Keller Used Syntactic Semantics to Escape from a Chinese Room*, in «Minds and Machines», 16(4), 381-436.

Reynolds R.I. (1988), *A Psychological Definition of Illusion*, in «Philosophical Psychology», 1(2), 217-223.

Rheingold H. (1991), *Virtual Reality: Exploring the Brave New Technologies*, New York, Simon & Schuster.

Ring M., Orseau L. (2011), *Delusion, Survival, and Intelligent Agents*, Paper presented at the International Conference on Artificial General Intelligence.

Robinson A.E., Hammon P.S., de Sa V.R. (2007), *Explaining Brightness Illusions Using Spatial Filtering and Local Response Normalization*, in «Vision Research», 47(12), 1631-1644.

Robinson J.O. (2013), *The Psychology of Visual Illusion*, Chelmsford, Courier Corporation.

Ropar D., Mitchell P. (1999), *Are Individuals with Autism and Asperger's Syndrome Susceptible to Visual Illusions?*, in «The Journal of Child Psychology and Psychiatry and Allied Disciplines», 40(8), 1283-1293.

Rupert R.D. (2004), *Challenges to the Hypothesis of Extended Cognition*, in «The Journal of Philosophy», 101(8), 389-428.

Schneider S., Turner E. (2017), *Is Anyone Home? A Way to Find Out If AI Has Become Self-Aware*, in «Scientific American», July 19,.

Schwarting M., Burton T., Yampolskiy R. (2015), *On the Obfuscation of Image Sensor Fingerprints*, Paper presented at the Information and Computer Technology (GOCICT), 2015 Annual Global Online Conference on.

Schweizer P. (2012), *Could There Be a Turing Test for Qualia?*, in *Revisiting Turing and His Test: Comprehensiveness, Qualia, and the Real World*, 41.

Seckel A. (2004), *Masters of Deception: Escher, Dalì and the artists of optical illusion*, New York, Sterling Publishing.

Segall M.H., Campbell D.T., Herskovits M.J. (1963), *Cultural Differences in the Perception of Geometric Illusions*, in «Science», 139(3556), 769-771.

Shibata K., Kurizaki S. (2012), *Emergence of Color Constancy Illusion Through Reinforcement Learning with a Neural Network*, Paper presented at the IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL).

Slater M., Spanlang B., Sanchez-Vives M.V., Blanke O. (2010), *First Person Experience of Body Transfer in Virtual Reality*, in «PLoS ONE», 5(5), e10564.

Soares N., Fallenstein B., Armstrong S., Yudkowsky E. (2015), *Corrigibility,* Paper presented at the Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, Texas, January 25-30.

Su J., Vargas D.V., Kouichi S. (2017), *One Pixel Attack for Fooling Deep Neural Networks*, arXiv preprint arXiv:1710.08864.

Sun J.T. (1924), *Psychology in Primitive Buddhism*, in «The Psychoanalytic Review» (1913-1957), 11, 39.

Suzuki K., Roseboom W., Schwartzman D.J., Seth A.K. (2017), *The Hallucination Machine: A Deep-Dream VR Platform for Studying the Phenomenology of Visual Hallucinations*, bioRxiv, 213751.

Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. (2013), *Intriguing Properties of Neural Networks*, arXiv preprint arXiv:1312.6199.

Tartaglia J. (2016), *What Is at Stake in Illusionism?*, in «Journal of Consciousness Studies», 23(11-12), 236-255.

Thaler S. (1993), *4-2-4 Encoder Death*, Paper presented at the Proceedings of the World Congress on Neural Networks.

Thaler S.L. (1995a), *Death of a Gedanken Creature*, in «Journal of Near Death Studies», 13, 149.

Thaler S. L. (1995b), *«Virtual Input» Phenomena Within the Death of a Simple Pattern Associator*, «Neural Networks», 8(1), 55-65.

Todrank J., Bartoshuk L.M. (1991), *A Taste Illusion: Taste Sensation Localized by Touch*, in «Physiology & Behavior», 50(5), 1027-1031.

Tong F., Engel S.A. (2001), *Interocular Rivalry Revealed in the Human Cortical Blind-Spot Representation,* in «Nature», 411(6834), 195-199.

Torrance S. (2012), *Super-Intelligence and (Super-) Consciousness*, in «International Journal of Machine Consciousness», 4(02), 483-501.

Trevarthen C. (2011), *What Is It Like to Be a Person Who Knows Nothing? Defining the Active Intersubjective Mind of a Newborn Human Being*, in «Infant and Child Development», 20(1), 119-135.

Tudusciuc O., Nieder A. (2010), *Comparison of Length Judgments and the Müller-Lyer Illusion in Monkeys and Humans*, in «Experimental Brain Research», 207(3-4), 221-231.

Tulving E., Schacter D.L. (1990), *Priming and Human Memory Systems*, in «Science», 247(4940), 301-306.

Turing A. (1950), *Computing Machinery and Intelligence*, in «Mind», 59(236), 433-460.

Tye M. (1999), *Phenomenal Consciousness: The Explanatory Gap as a Cognitive Illusion*, in «Mind», 108(432), 705-725.

Velan H., Frost R. (2007), *Cambridge University Versus Hebrew University: The Impact of Letter Transposition on Reading English and Hebrew*, in «Psychonomic Bulletin & Review», 14(5), 913-918.

Vokey J.R., Read J.D. (1985), *Subliminal Messages: Between the Devil and the Media*, in «American Psychologist», 40(11), 1231.

Wallaeh H., Kravitz J.H. (1965), *The Measurement of the Constancy of Visual Direction and of Its Adaptation*, in «Psychonomic Science», 2(1-12), 217-218.

Walter W.G., Dovey V., Shipton H. (1946), *Analysis of the Electrical Response of the Human Cortex to Photic Stimulation*, in «Nature», 158(4016), 540-541.

Wegner D.M. (1994), *Ironic Processes of Mental Control*, in «Psychological Review», 101(1), 34.

Wells A. (2012), *The Literate Mind: A Study of Its Scope and Limitations*, New York, Palgrave Macmillan.

Wickler W. (1968), *Mimicry in Plants and Animals,* London, Weidenfeld & Nicolson.

Yamins D.L., DiCarlo J.J. (2016), *Using Goal-Driven Deep Learning Models to Understand Sensory Cortex*, in «Nature Neuroscience», 19(3), 356-365.

Yampolskiy R. (2007), *Graphical CAPTCHA Embedded in Cards, Western New York Image Processing Workshop (WNYIPW)-IEEE Signal Processing Society*, Rochester, NY, September.

Yampolskiy R. (2013), *Turing Test as a Defining Feature of AI-Completeness*, in X.-S. Yang (ed.), *Artificial Intelligence, Evolutionary Computing and Metaheuristics*, Berlin, Springer, pp. 3-17.

Yampolskiy R., Cho G., Rosenthal R., Gavrilova M. (2012), *Experiments in Artimetrics: Avatar Face Recognition*, in «Transactions on Computational Science», XVI, 77-94.

Yampolskiy R., Fox J. (2013), *Safety Engineering for Artificial General Intelligence*, in «Topoi», 32(2), 217-226.

Yampolskiy R. V. (2012), *AI-Complete CAPTCHAs as Zero Knowledge Proofs of Access to an Artificially Intelligent System*, in «ISRN Artificial Intelligence», ID 271878.

Yampolskiy R.V. (2013a), *Artificial Intelligence Safety Egineering: Why Machine Ethics is a Wrong Approach*, in «Philosophy and Theory of Artificial Intelligence», 389-396.

Yampolskiy R.V. (2013b), *Attempts to Attribute Moral Agency to Intelligent Machines Are Misguided*, Paper presented at the Proceedings of Annual Meeting of the International Association for Computing and Philosophy, University of Maryland at College Park, MD.

Yampolskiy R.V. (2013c), *Efficiency Theory: A Unifying Theory for Information, Computation and Intelligence*, in «Journal of Discrete Mathematical Sciences & Cryptography», 16(4-5), 259-277.

Yampolskiy R.V. (2014), *Utility Function Security in Artificially Intelligent Agents*, in «Journal of Experimental & Theoretical Artificial Intelligence», 26(3), 373-389.

Yampolskiy R. V. (2015a), *Artificial superintelligence: a futuristic approach*, Boca Raton (FL), CRC Press.

Yampolskiy R.V. (2015b), *The Space of Possible Mind Designs*, Paper presented at the International Conference on Artificial General Intelligence.

Yampolskiy R.V. (2016a), *On the Origin of Synthetic Life: Attribution of Output to a Particular Algorithm*, in «Physica Scripta», 92(1), 013002.

Yampolskiy R.V. (2016b), *Taxonomy of Pathways to Dangerous Artificial Intelligence*, Paper presented at the AAAI Workshop: AI, Ethics, and Society.

Yampolskiy R.V. (2017), *Future Jobs – The Universe Designer,* Paper presented at the Circus Street, Available at: https://blog.circusstreet.com/future-jobs-the-universe-designer/.

Yampolskiy R.V. (2017), *What Are the Ultimate Limits to Computational Techniques: Verifier Theory and Unverifiability*, in «Physica Scripta», 92(9), 093001.

Yampolskiy R.V. (2018), *Artificial Consciousness: An Illusionary Solution to the Hard Problem*, in «Reti saperi linguaggi. Italian Journal of Cognitive Sciences», 2.

Yampolskiy R.V. (2012), *AI-Complete, AI-Hard, or AI-Easy – Classification of Problems in AI*, Paper presented at the The 23rd Midwest Artificial Intelligence and Cognitive Science Conference, Cincinnati, OH, USA, April 21-22.

Yampolskiy R. V., Fox J. (2012), *Artificial Intelligence and the Human Mental Model*, in A. Eden, J. Moor, J. Soraker, E. Steinhart (eds.), *In the Singularity Hypothesis: a Scientific and Philosophical Assessment*, Berlin, Springer.

Yampolskiy R.V., Gavrilova M.L. (2012), *Artimetrics: Biometrics for Artificial Entities*, in «Robotics & Automation Magazine, IEEE», 19(4), 48-58.

Yampolskiy R.V., Govindaraju V. (2007), *Embedded Non-Interactive Continuous Bot Detection*, in «ACM Computers in Entertainment», 5(4), 1-11.

Yampolskiy R.V., Rebolledo-Mendez J.D., Hindi M.M. (2014), *Password Protected Visual Cryptography via Cellular Automaton Rule 30*, in *Transactions on Data Hiding and Multimedia Security IX*, Berlin, Springer, pp. 57-67.

Yampolskiy R.V., Spellchecker M. (2016), *Artificial Intelligence Safety and Cybersecurity: A Timeline of AI Failures*, arXiv preprint arXiv:1610.07997.

Yudkowsky E. (2015), *Rationality: From AI to Zombies*, Berkeley, MIRI.

Zadra A., Donderi D. (2000), *Nightmares and Bad Dreams: Their Prevalence and Relationship to Well-Being*, in «Journal of Abnormal Psychology», 109(2), 273.

Zeman A. (2015). *Computational modelling of visual illusions*, PhD Thesis, Macquarie University.

Zeman A., Brooks K.R., Ghebreab S. (2015), *An Exponential Filter Model Predicts Lightness Illusions*, in «Frontiers in Human Neuroscience», 9.

Zeman A., Dewar M., Della Sala S. (2015), *Lives Without Imagery – Congenital Aphantasia*, in «Cortex», 73(Supplement C), 378-380.

Zeman A., Obst O., Brooks K.R. (2014), *Complex Cells Decrease Errors for the Müller-Lyer Illusion in a Model of the Visual Ventral Stream*, in «Frontiers in Computational Neuroscience», 8.

Zeman A., Obst O., Brook, K.R., Rich A.N. (2013), *The Müller-Lyer Illusion in a Computational Model of Biological Object Recognition*, in «PLOS ONE», 8(2), e56126.