Authors: 1. Ojo Akinyemi,
Lecturer III
Department of Social Sciences,
Rufus Giwa Polytechnics, Owo.
Ondo State.
akyem200@gmail.com
07038374883

2. Ilepe Johnson Akintunde
Pincipal Lecturer
Department of Social Sciences,
Rufus Giwa Polytechnics, Owo.
Ondo State.
johnsonilepe@gmail.com
08033894364

A CRITICAL APPRAISAL OF ARTIFICIAL INTELLIGENCE

ABSTRACT

The study critically examined the analytical implication imbued in Artificial Intelligence, which denotes that human intelligence can be simulated by electronic machines such as Computer. Human claims to be uniquely rational have come under pressure from investigators in the rapidly expanding field of artificial intelligence (AI). It is customary, in this context, to distinguish between ‘strong and weak’ Artificial Intelligence, proponents of the latter maintain simply that aspects of intelligent human behaviour can be usefully simulated or modeled by appropriately programmed computers, whereas, chauvinists of the former hold tenaciously that in virtue of executing a suitably written programme, a machine could literally be said to think and reason.

Clearly stating, a strong Artificial Intelligence whose claims are philosophically controversial and whose credentials we must therefore investigate. Using a critical evaluated method and a review of existing literature, the study examined the position of strong artificial intelligence, which had generated a lot of discordance and dissonance among scholars in philosophy and in some other related fields. As a corollary, diverse theories had emerged either in its defense or relegating it to a status of weak artificial intelligence. However, none of these theories is even adequate to solve the age-long problems. Consequently, the study evaluated an America Philosopher-John Roggers Searle's Chinese Room argument who was a patriot of weak Artificial Intelligence. The study unmasked the fact that artificial intelligence is not homonymous to human intelligence especially in term of rationality, consciousness, self-awareness, emotions and much else, upon which some Philosophers, Psychologists and Scientists based their arguments. The study argued that simulation or modeling of intelligent human behaviour by electronically programmed computer only demean and degrade the natural paraphernalia and capacity of human intelligence. Following from the foregoing, the study showed that it was nearer to intellectual barrenness to conclude on the basis of ignorance of something, as the partisans of strong artificial intelligence had done in the defense of their position. Considering the logical sequence of the arguments pursued, the study simply recommended that strong artificial intelligence sits impotent in the face of critical intellectual contest and should be adequately retooled in order to be ancillary to frontier of knowledge.

KEYWORDS:
Robots, Intelligence, Computer, Consciousness, Technology

INTRODUCTION

The term intelligence refers to the ability to acquire and apply different skills and knowledge to solve a given problem. In addition, intelligence is also concerned with the use of general mental capability to reason and learn various situations (R. Feverstein). Intelligence is intertwined with multifaceted cognitive functions such as; language, attention, planning, memory, perception and so on. Intelligence involves both Human and Artificial Intelligence. In this case, critical human intelligence is concerned with solving problems, reasoning and learning (I. S. Cottfredson).

The information technology revolution has led to ambitious pronouncements by researchers on the field of artificial intelligence, some of whom maintain that suitably programmed computers can literally be said to engage in processes of thought and reasoning. Ironically enough, some empirical psychologists have begun to challenge our own human pretensions to be able to think rationally, We are thus left contemplating the strange proposition that machines of our own devising may soon be deemed rational than their human creators. Whether we can make coherent sense of such a suggestion is one issue that we may hope to resolve in the course of this work. But to be in a position to do so, we need to examine more closely, the nature and basis of some of the surprising claims being made by investigators in the field of artificial intelligence. Some of the questions that we should consider are the following, Could an electronic machine literarily be said to engage in processes of thought and reasoning, simply by virtue of executing a suitably formulated computer programme? Or can we at best talk of computers as simulating rational thought processes for the purpose of weather forecasting? Would a genuinely intelligent machine have to possess a brain? Would a physical configuration somewhat similar to that of a human brain? Will it need to be conscious, to be able to learn by experience and capable of interacting intentionally with its physical and social environment? How far are the intelligence and rationality a matter of possessing what might be called ‘commonsense’? Could it be captured in a computer programme? What is the Cartesian view on mindedness of electronic machines? Without more ado, let us now start x-raying the field of artificial intelligence, so as to proffer possible answers to the above and related questions.

Could A Computer Belief?

Obviously, it is imperative to focus on the question of whether computers could have minds. Could a computer believe? At first sight, the answer seems obvious. No, computers are made of silicon chips rather than neuroproteins; they house dry hardware rather than brains. By contrast, humans and animals possess brains. One is tempted to say, for example, as the American philosopher John Rogers Searle, says 'neurons firing in specific neural architectures' simply put, brains are necessary for belief as well as for thought and experience more generally (Rusell&Norvig, 2009). Since computers are brainless, they are mindless. However, partisans of the idea that computers could believe challenge the position of Searle as mere begging the question when he insists that brains are necessary for belief, Searle assumes what needs adequate proof. One counter argument is that; if brains are unnecessary for intelligent behavior and intelligent behavior is sufficient for belief, and computer behave sufficiently intelligently, then even though computers are brainless, they believe. More generally, they possess minds of their own (Lighthill, 1973).

David Hume posits that the mark of mindedness in an animal is the capacity for intelligent behavior, not perhaps of the flatworm but of the bird or dog. Now it is common place knowledge that computers behave intelligently; they win chess games, perform numerical calculations and guide missiles. On this behavioural way of thinking, if computer behavior is sufficiently intelligent, computers have minds. So misgivings about Searle's deference to brains arise. Could a computer believe? Even though its brainless tempts saying no, it's behaviours prompts saying yes.

What are the prospects for resolving the debase between brain devotees like Searle and more general attributors, between those who contend that brain is essential and those who fix on behavior?

Imagine that a creative from another planet lands in your backyard and tells you that it is going to enroll for courses at Princeton University. Within a little period of time, it is awarded a Ph.D in Comparative Literature! What should we say? Should we say that brains are since qua non for belief or that they are inessential? In this case, most of us would think it obvious what we should say; we should say that brains are unnecessary. It has a mind of its own even if it lacks a brain. Though, this is fictitious example; but it shows that Searle's

particular and initial insistence on brains is untenable and unreasonable. The basic idea is this; if an individual's behavior is sufficiently intelligent, if it is, for instance, relevantly similar to human behavior of the sort which in us warrants the attribution of belief, then there is good reason to say that it believes, even if it lacks a brain.

ALAN TURING TEST

Alan Turing (1912 – 1954), a logician per excellence and father of a lot of the modern theory of computers and computation, had another behavioural criterion for whether computers are minded. For him, not chess, but conversation, he proposed, in a paper which appeared in the Journal Mind in 1950, that if a computer could communicate with a human being in such a way that the human being could not tell the difference between conversing with the computer and conversing with another human being, then it would be rather absurd and arbitrary to deny that computer is a believer simply because it is a computer. Turing argued that the capacity of linguistic conversation is a good, strong for computer belief. Wouldn't it be impressive if a computer could speak a language, say Chinese, and engage in real dialogue? Its conversational range might manifest a thick and elaborate belief network, a stock of attitudes visible through its 'speech', concerning such topics as chess, monopoly, games of all sorts-on boards, in love, politics and life. Who could deny intrinsic intentionality of an intelligent conversationalist, even though it was brainless?

John Rogers Searle has constructed an imaginative 'thought experiment', also known as Chinese room argument to challenge the validity of Turing's proposal that computer conversation would demonstrate computer belief. Thought experiments are important analytical devices in the history of philosophy. Some of the examples of these are;

The case of the League Alien, Descartes' Evil Demon argument and Plato's Ring of Gyges. A thought experiment consists in hypothesizing certain imaginary assumptions and extracting insights from those assumptions. Plato, for instance, supposes that a person possesses a magical ring the rubbing of which enables him to become invisible and to commit undetected immoralities and misdeeds. By this device, Plato explores the view, which he himself abjures or disclaim, that we follow moral rules merely because we fear punishment and not because we respect morality for itself. The Ring of Gyges is a thought experiment in human moral psychology.

The Chinese Room Argument

Searle's Chinese room is a thought experiment in computer conversational psychology. It goes roughly as follows;

Imagine that a man is locked in an enclosed room with one postal slot through which symbols in a language, say Chinese, are sometimes sent in and one exit slot through which symbols can be sent out. In the centre of the room are several baskets containing various sorts of Chinese symbols. Then, if the person in the room was fluent in his native language, say English, but ignorant of the language of the symbols, and had a book of instructions written in his native language which told him to send certain Chinese symbols out when certain other symbols were sent in, it would seems to outside observers as if the person in the room understood Chinese. However, to him Chinese writing looks like meaningless squiggles, nor does the book of instructions decode their meaning, for the book, we may imagine, identifies symbols entirely by their physical shapes and does not require that the man understand any of them. The book is not a dictionary but a set of rules which tells which symbols to send back out of the exit slot. Take a squiggle-squaggle from basket number one and place it next to a squaggle-squiggle from basket number two; then bring both symbols to exit slot. In following such instructions, the person seems to outside observers to answer Chinese questions and comments received at the entry slot with responses delivered at the exit slot. For example, someone outside might hand him symbols that unknown to him mean, 'Who is your favourite philosopher?' and he might after following the rule book hand back symbols that, also unknown to him, mean, 'My favourite philosopher is Donald Davidson, though I admire Martin Buber a lot'.

While unapparent to outside observers, Searle contends, it remains certain that the person in the room does not understand Chinese. He applies instructions, and he understands those instructions, but the sequences of Chinese symbols are gibberish to him. Also clear, contends Searle, is that the entire system of the room plus the person and its contents is ignorant of Chinese as well. Nothing here grasps Chinese, except those sending and receiving messages, and those who composed the instruction book. No conversational response of man or system has an intrinsic. Intentionality, or is about anything, or means anything, except as it is viewed by outside observers interacting with it.

Searle contends that the Chinese room argument hits the mark of undermining computer intrinsic. Intentionality, belief and understanding, for the setup contains everything relevant to a computer. The person in the room, functioning like the central processor of a computer, manipulates physical symbols (in computer jargon 'syntax'); the instruction book is the 'computer program' for manipulating the symbols. The baskets full of symbols are the 'data base'; the small bunches that are handed in are the 'input' and the bunches then handed out are the 'output'. According to Searle the impact of example is as follows; if person in the room (or system) does not possess intrinsic Intentionality (does not understand Chinese) solely on the basis of following a computer program for understanding Chinese, then neither does any computer on the basis of running its program possess intrinsic Intentionality. Manipulating symbols in accordance with rules, which is the essence of computer program operation and transpires in the Chinese Room, produces mere as-if Intentionality, as if belief, as if Chinese understanding.

Given Searle's conclusion, there cannot be any dispute about whether passing the Turing test is sufficient for computer belief. For it is not. Far from being the product of intrinsic Intentionality, computer conversation by itself reveals nothing more than as if Intentionality and as if belief.

Searle first presented the Chinese room in the pages of *Behavioural and Brain Sciences* in 1980, and it has since appeared and has been refined in several of his publications. It also has been subjected to a flood of comment, discussion and criticism, including 26 commentaries to the original article (Williams).

Even to critics much of what Searle says seems both right and important. He is clearly correct in his contention that the example shows that someone or something can appear to understand without genuinely understanding; or appear to converse without really conversing; or appear to believe without truly believing. In life, we may simulate characteristics or abilities we really do not possess. Behavior can mask ignorance. Also, Searle is on to something significant in his thesis that as if Intentionality differs from intrinsic Intentionality. Few would deny that people differ from water flowing downhill. The distinction between as if and intrinsic Intentionality may mark the psychological divide which separates water from persons and non-believers from believers. However, we can agree that these are good points without agreeing

with Searle that computers could not believe. Many philosophers continue to contend that computers in one way or another, could believe.

The most popular anti-Searle, pro-computer belief counter-argument attacks Searle at just this point: it urges that various other possibilities, in addition to the Chinese room, are relevant to determining whether a computer could believe. Therefore, the Chinese room does not undermine computer intrinsic Intentionality.

DESCARTES' REMARKS ON MIND AND ARTIFICIAL INTELLIGENCE

So far, we have x-rayed artificial intelligence and its presuppositions of mind. In the Cartesian scheme of mind, there is no place for computationality because the thought act is due to the subjective thinking thing, which is the self. Again, this subjective thinking thing or the self is that which "doubts, understands, affirms, denies, is willing, is unwilling and also imagines and has sensory perceptions" (Descartes, 1984). The existence of the thinking thing is the same as the existence of the subjective thinking thing, because it is the subject, who thinks. All these subjective activities are noncomputational because the subjective activity is the first person perspective. The mental processes, for Descartes, are intentional and are the free acts of the thinking subject. Therefore, this subjective attitude of mind cannot be mapped mechanically in an algorithmic system.

Descartes' concept of 'I think' presupposes subjective experience, because it is 'I' who experiences the world. Likewise, Descartes' notion of 'I' negates the notion of computationality of the mind. The essence of mind is thought and the acts of thoughts are identified with acts of consciousness. Therefore, it follows that cognitive acts are conscious acts, but not computational or mechanical acts. Thus for Descartes, one of the most important aspects of cognitive states and processes is their phenomenality because of our judgments, understanding, etc. that can be defined and explained only in relation to consciousness and not in relation to computationality. We can only find computationality in machines and not in the mind, which wills, understands and judges. Descartes' dictum, "I think, therefore, I am" (Descartes, 1984). Not only establishes the existence of the self which thinks and acts but also its freedom from mechanistic laws to which the human body is subjected to.

Moreover, when Descartes makes the distinction between mind and body, he did not claim that the idea of the mind is that of a ghost, although he did say that the idea of the body

is that of a machine. Following this, Ryle in his book, 'The Concept of Mind' says that Descartes' distinction between mind and body is a myth. He argues, "I shall often speak of it, with deliberate abusiveness, as 'the dogma of the ghost in the machine'. I hope to prove that it is entirely false, and false not in detail but in principle" (Ryle, 1985). According to Ryle, Descartes' distinction between mind and body commits a categorymistake (Ryle, 1985).

As Ryle said, "my destructive purpose is to show that a family of radical category mistake is the source of the double-life theory. The representation of a person as a ghost mysteriously ensconced in a machine derived from this argument, because, as is true, a person's thinking, feeling and purposive doing cannot be described solely in the idioms of physics, chemistry and physiology. Therefore they must be described in counterpart idioms. As the human body is a complex organized unit, so the human mind must be another complex organized unit, though one made of a different sort of stuff and with a different sort of structure. Likewise, again, as the human body, like any other parcel of matter, is a field of causes and effects, so the mind must be another field of causes and effects, though not (Heaven be praised) mechanical causes and effects" (Ryle, 1985).

In Ryle's understanding of mind, mind becomes as much mechanical as the body and is therefore non-different from the body. However, Descartes refutes the mechanistic reading of mind. As we have seen, Descartes is a dualist, rather than a mentalist. Descartes' argument for the mind, which is distinct from the body, needs to be understood as an argument for the logical possibility of their separate existence and not for the fact that they exist independent of each other. The separability argument is as follows: "First, I know that everything, which clearly and distinctly understands is capable of being created by God so as to correspond exactly with my understanding of it. Hence, the fact that I can clearly and distinctly understand one thing apart from another is enough to make certain that two things are distinct, since they are capable of being separated at least by God. The question of what kind of power is required to bring about such a separation does not affect the judgment that the two things are distinct. Thus, simply by knowing that I exist and seeing at the same time that absolutely nothing else belongs to my nature or essence except that I am a thinking thing, I can infer correctly that my essence consists solely in the fact that I am a thinking thing. It is true that I may have (or, to anticipate, that I certainly have) a body that is very closely joined to me, nevertheless, on the

one hand, I have a clear and distinct idea of myself, in so far as I am simply a thinking, nonextended thing; and on the other hand, I have a distinct idea of the body, in so far as this is simply an extended non-thinking thing. Accordingly, it is certain that I am really distinct from my body and can exist without it" (Descartes, 1984).

Descartes has already proved in the 'Second Meditation' the existence of a thinking being that has a clear and distinct perception of the mind as a thinking, non-extended thing. This is a proof of the non-mechanical mind which is different from the body subject to mechanical laws. Similarly, in the 'Fifth Meditation', he has shown that he has a clear and distinct idea of a body as extended and a non-thinking substance. This is to suggest that the mechanically existing body is ontologically distinct from the non-computational mind.

The afore-described distinction between mind and body supposes that there is no 'ghost' in the human body or 'ghost in the machine'. However, Descartes did not admit the existence of ghost in the machine. Had Descartes admitted that there was a ghost in the human body, then the mind itself would become computational, and there would be no necessary distinction between the mind and the body, because the ghost itself is a body; but Descartes admits the distinction between mind and body and this shows that the mind is non-computational. It is mind, which has the capacity of intelligence and understanding. The Cartesian way of understanding the concept of intelligence is anti-physicalist and antibehaviourist and hence is anti-mechanical. The human mind is beyond the sphere of computationality, because the human mind has innate ideas, which are embedded as the innate dispositions of the human mind. These ideas are a priori in the human mind and are the basic in-born propensities. Descartes observes, "my understanding of what a thing is, what truth is and what thought is, seems to be derived simply from my own nature, but my hearing a noise, as I do now, or seeing the sun, or feeling the fire, comes from the thing which is located outside of me, or so I have hitherto judged. Lastly, sirens, hippogriffs and the likes are my own invention" (Descartes, 1984).

The afore-said observation of Descartes shows that innate ideas are not produced in us by the senses. If the ideas were conveyed to us by the senses like heat, sound, etc., we would not have to refer to anything outside ourselves, because they too would be innate. For Descartes, "the ideas of pain, colours, sounds and the likes must be all the more innate, if, on

the occasion of certain corporeal motions, our mind is to be capable of representing them to itself, for there is no similarity between these ideas and the corporeal motions." (Descartes, 1985). Here, it follows that there is a distinction between innate and adventitious ideas and that innate ideas are universal ideas, whereas adventitious ideas are particular ideas. As such, Descartes points out that hearing a noise, seeing the scene and feeling the fire are all particular ideas. (Descartes, 1984). Again, it must be noted that the perception of the particular is not possible without the universal. Innate universal ideas are a necessary requirement for the cognition of the particular objects in the world.

Descartes holds that thinking cannot be explained mechanically. His argument that brutes cannot think is equivalent to an argument that machines cannot think. He thinks that no machine could have the capacity of using the linguistic and other signs to express thoughts and to give appropriate responses to meaningful speech, and the capacity to act intelligently or rationally in all sorts of situations (Descartes, 1985). But what is so special about human language use and what does it show that the behaviour of any mechanism fails to show? A machine could be construed to utter words corresponding to bodily change in its origin, but could never use spoken words or other signs that are composed as we do to declare our thoughts to others, because "It is not conceivable that the machine should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as the dualist of men can do. Secondly, even though such machines might do some things as well as we do them, or perhaps even better, they would inevitably fail in others, which could reveal that they were acting not through understanding but only from the disposition of their organs. For the fact that reason is a universal instrument which can be used in all kind of situations, these organs need some particular disposition for each particular action, hence it is morally impossible to have enough different ones in a machine to make it act in all contingencies of life in the way in which our reason makes us act" ((Descartes, 1985).

What Descartes is drawing attention to here is firstly, no machine could have the capacity to use linguistic and other signs to express thoughts and give appropriate responses to meaningful speech, and secondly, machine would not have the capacity to act intelligently in all sorts of situations. Here, animal communication have not offered counter evidence to

Descartes' assumption that human language is based on an entirely distinct principle, nor has modern linguistic philosophers dealt with their observations in serious way. For Chomsky, the main lessons to be learnt from the 'Cartesian' tradition in linguistic are the idea of an innate, universal grammar and the idea that the study of the structure of this argument will reveal the structure of thought or mind.

Descartes' argument that brutes or machines cannot think in the light of the general question of what makes an utterance or a symbolic structure meaningful is noteworthy. The kind of automatic, rule governed computation or symbol processing that a turing machine instantiates and that can be performed by electronic computers would not count as thinking in Descartes' sense, nor would the mechanical operations of a computer or robot, no matter how ingenious or intelligent, count as rational behaviour as he understands it. Not only did such a view makes thinking too narrow, it is based on precisely the kind of category mistake that Ryle attributes to the Cartesians which have been discussed earlier.

Descartes is not a reductionist as he thinks that mind cannot be reduced to anything else and it must have an autonomous existence alongside the existence of the material body. The 'I think' of the mental reality does not deny the 'I exist' character in the world, rather it is an affirmation of it. In that sense, we cannot say that Descartes has subjectivized the mental world and thus made it into a private world. He made every effort to keep an objective constraint on the subjective mind and thus forestalled all skeptical questions about the existence of other minds. (Pradhan, 2001). This is because Cartesian doctrine of the mind and its inner experience do assume that we know other minds as much as we know our own. That is the reason why Descartes called the 'I think' the absolute basis of all our knowledge-claims about others and the external world. Thus, the self or mind is irreducible/not explainable in terms of the body or machines whether of Descartes or another's. In view of this, we can say that the Cartesian philosophy of the mind is not based on a mistake and as such, it has shown the right way to the understanding of the mind.

Of course Descartes would not have accepted the idea of mechanical or computational artificial intelligence, because he may still be considered an important forerunner of cognitive and computational view of the mind. The essence of the mind is rational thinking and rational thought or cognition can be studied independently of other phenomena, like sensation and

emotions, in that Descartes stated that the body is dependent on mental phenomena, to which the mind is referred to as consciousness. Although Descartes did not identify mental thought with consciousness, emotions, awareness, etc., he regarded that all these are conditions of thought. While arguing the existence of mind, descartes talks about the mind acting in some particular location in the brain. As such, this is comparable to contemporary literal talk about mental processes as computational activity in the brain. Moreover, Descartes would not have accepted the mechanical application of rules on syntactic structures as a sufficient condition for rational symbol manipulation. The kind of automatic, rule governed computation or symbol processing that a turning machine instantiates and that can be performed by electronic computers would not count as thinking from the Cartesian point of view. Therefore, Cartesian thinking is neither reducible to a narrowly understood rational capacity nor to consciousness because he clearly mentions that consciousness is a necessary condition for thought.

Weak and Strong Artificial Intelligence

When defining the capacity of Artificial Intelligence, this is frequently categorised in terms of weak or strong Artificial Intelligence. Weak Artificial Intelligence (narrow Artificial Intelligence) is one intended to reproduce an observed behaviour as accurately as possible. It can carry out a task for which they have been precisiontrained. Such Artificial Intelligence systems can become extremely efficient in their own field but lack generalisational ability. Most existing intelligent systems that use machine learning, pattern recognition, data mining or natural language processing are examples of weak Artificial Intelligence. Intelligent systems, powered with weak Artificial Intelligence include recommender systems, spam filters, self-driving cars, and industrial robots. Strong Artificial Intelligence is usually described as an intelligent system endowed with real consciousness and is able to think and reason in the same way as a human being. A strong Artificial Intelligence can, not only assimilate information like a weak AI, but also modify its own functioning, i.e. is able to autonomously reprogram the Artificial Intelligence to perform general intelligent tasks. These processes are regulated by human-like cognitive abilities including consciousness, sentience, sapience and self-awareness. Efforts intending to generate a strong Artificial Intelligence have focused on whole brain simulations, however this approach has received criticism, as intelligence cannot be simply explained as a biological process emanating from a single organ

but is a complex coalescence of effects and interactions between the intelligent being and its environment, encompassing a series of diverse ways via interlinked biological process.

Questioning the Impact of Artificial Intelligence

Given the exponential rise of interest in Artificial Intelligence, experts have called for major studies on the impact of Artificial Intelligence on our society, not only in technological but also in legal, ethical and socioeconomic areas. This response also includes the speculation that autonomous super artificial intelligence may one day supersede the cognitive capabilities of humans. This future scenario is usually known in Artificial Intelligence forums as the "Artificial Intelligence singularity" [Spinrad, 2017]. This is commonly defined as the ability of machines to build better machines by themselves. This futuristic scenario has been questioned and is received with scepticism by many experts. Today's Artificial Intelligence researchers are more focused on developing systems that are very good at tasks in a narrow range of applications. This focus is at odds with the idea of the pursuit of a super generic Artificial Intelligence system that could mimic all different cognitive abilities related to human intelligence such as self-awareness and emotional knowledge. In addition to this debate, about Artificial Intelligence development and the status of our hegemony as the most intelligent species on the planet, further societal concerns have been raised. For example, the Artificial Intelligence 100 (One Hundred Year Study on Artificial Intelligence) a committee led by Stanford University, defined 18 topics of importance for Artificial Intelligence [Horvitz, 2014]. Although these are not exhaustive nor definitive, it sets forth the range of topics that need to be studied, for the potential impact of Artificial Intelligence and stresses that there are a number of concerns to be addressed. Many similar assessments have been performed and they each outline similar concerns related to the wider adoption of Artificial Intelligence I technology.

The 18 topics covered by the Artificial Intelligence 100

- **Technical trends and surprises**: This topic aims at forecasting the future advances and competencies of Artificial Intelligence technologies in the near future. Observatories of the trend and impact of Artificial Intelligence should be created,

helping to plan the setting of Artificial Intelligence in specific sectors, and preparing the necessary regulation to smooth its introduction.

- **Key opportunities for Artificial Intelligence:** How advances in Artificial Intelligence can help to transform the quality of societal services such as health, education, management and government, covering not just the economic benefits but also the social advantages and impact.
- **Delays with translating Artificial Intelligence advances into real-world values**: The pace of translating AI into real world applications is currently driven by potential economic prospects [Lohr, 2012]. It is necessary to take measures to foster a rapid translation of those potential applications of Artificial Intelligence that can improve or solve a critical need of our society, such as those that can save lives or greatly improve the organisation of social services, even though their economic exploitation is not yet assured.
- **Privacy and machine intelligence**: Personal data and privacy is a major issue to consider and it is important to envisage and prepare the regulatory, legal and policy frameworks related to the sharing of personal data in developing Artificial Intelligence systems.
- **Democracy and freedom**: In addition to privacy, ethical questions with respect to the stealth use of Artificial Intelligence for unscrupulous applications must be considered. The use of Artificial Intelligence should not be at the expense of limiting or influencing the democracy and the freedom of people.
- **Law:** This considers the implications of relevant laws and regulations. First, to identify which aspects of Artificial Intelligence require legal assessment and what actions should be undertaken to ensure law enforcement for Artificial Intelligence services. It should also provide frameworks and guidelines about how to adhere to the approved laws and policies.
- **Ethics:** By the time Artificial Intelligence is deployed into real world applications there are ethical concerns referring to their interaction with the world. What uses of Artificial Intelligence should be considered unethical? How should this be disclosed?

- **Economics**: The economic implications of Artificial Intelligence on jobs should be monitored and forecasted such that policies can be implemented to direct our future generation into jobs that will not be soon overtaken by machines. The use of sophisticated Artificial Intelligence in the financial markets could potentially cause volatilities and it is necessary to assess the influence Artificial Intelligence systems may have on financial markets.
- **Artificial Intelligence and warfare**: Artificial Intelligence has been employed for military applications for more than a decade. Robot snipers and turrets have been developed for military purposes [Alston, 2011]. Intelligent weapons have increasing levels of autonomy and there is a need for developing new conventions and international agreements to define a set of secure boundaries of the use of Artificial Intelligence in weaponry and warfare.
- **Criminal uses of Artificial Intelligence**: Implementations of Artificial Intelligence into malware are becoming more sophisticated thus the chances of stealing personal information from infected devices are getting higher. Malware can be more difficult to detect as evasion techniques by computer viruses and worms may leverage highly sophisticated Artificial Intelligence techniques [Young & Yung, 1997; Kirat*et al.,* 2014]. Another example is the use of drones and their potential to fall into the hands of terrorists the consequence of which would be devastating.
- **Collaboration with machines:** Humans and robots need to work together and it is pertinent to envisage in which scenarios collaboration is critical and how to perform this collaboration safely. Accidents by robots working side by side with people had happened before [Bryant & Waters, 2015] and robotic and autonomous systems development should focus on not only enhanced task precision but in also being able to understand the environment and human intention.
- **Artificial Intelligence and human cognition**: Artificial Intelligence has the potential for enhancing human cognitive abilities. Some relevant research disciplines with this objective are sensor informatics and humancomputer interfaces. Apart from applications to rehabilitation and assisted living, they are also used in surgery [Andreu-Perez *et al.,* 2016] and air traffic control [Harrisson*et al.,* 2014]. Cortical implants are

increasingly used for controlling prosthesis, our memory and reasoning are increasingly relying on machines and the associated health, safety and ethical impacts must be addressed.

- **Safety and Autonomy:** For the safe operation of intelligent, autonomous systems, formal verification tools should be developed to assess their safety operation. Validation can be focused on the reasoning process and verifying whether the knowledge base of an intelligent system is correct [Gonzalez & Barr, 2000] and also making sure that the formulation of the intelligent behaviour will be within safety boundaries [Ratschan& She, 2005].
- **Loss of control of Artificial Intelligence systems**: The potential of Artificial Intelligence being independent from human control is a major concern. Studies should be promoted to address this concern both from the technological standpoint and the relevant framework for governing the responsible development of Artificial Intelligence.
- **Psychology of people and smart machines**: More research should be undertaken to obtain detailed knowledge about the opinions and concerns people have, in the wider usage of smart machines in societies. Additionally, in the design of intelligent systems, understanding people's preferences is important for improving their acceptability [Broadbent *et al.*, 2009; Smarr*et al.,* 2012].
- **Communication, understanding and outreach:** Communication and educational strategies must be developed to embrace Artificial Intelligence technologies in our society. These strategies must be formulated in ways that are understandable and accessible by non-experts and the general public.
- **Neuroscience and Artificial Intelligence:** Neuroscience and Artificial Intelligence can develop together. Neuroscience plays an important role for guiding research in Artificial Intelligence and with new advances in high performance computing, there are also new opportunities to study the brain through computational models and simulations in order to investigate new hypotheses [Reilly & Munakata2000].
- **Artificial Intelligence and philosophy of mind:** When Artificial Intelligence can experience a level of consciousness and self-awareness, there will be a need to

understand the inner world of the psychology of machines and their subjectivity of consciousness.

ETHICAL AND LEGAL QUESTIONS OF ARTIFICIAL INTELLIGENCE

Ethical Issues in Artificial Intelligence

Threat to Privacy

Data is the "fuel" of Artificial Intelligence and special attention needs to be paid to the information source and if privacy is breached. Protective and preventive technologies need to be developed against such threats. Although solutions to this issue may be unconnected to the Artificial Intelligence operation per se, it is the responsibility of Artificial Intelligence operators to make sure that data privacy is protected. Additionally, applications of Artificial Intelligence, which may compromise the rights to privacy, should be treated with special legislation that protects the individual.

Threats to Security and Weaponisation of Artificial Intelligence

With the proliferation of security risks, such as terrorism and regional conflicts, we are immersed in a global arms race that has developed a demand for new weapons powered by Artificial Intelligence, such as autonomous drones and missiles, as well as virtual bots and malicious software aimed at sophisticated espionage. This could lead to the possibility of an unprecedented war escalation. The danger of Artificial Intelligence is that we could potentially lose control over it. This has led associations and non-governmental organizations (NGOs) to carry out awareness-raising actions to contain the use of these military robots and potentially ban the use of such weapons.

The US military has just released a prospective document entitled 'Robotic and Autonomous Systems Strategy', which details its strategy for robotics and autonomous systems. The use of robotics and autonomous systems by the US military defines the following five objectives:

1. Increase knowledge abilities in operations' theatres,
2. Extenuate the amount of charge carried by the soldier;
3. Improve logistics capacity;
4. Enhance movement and maneuvering;

5. Boost the protection of forces. Although offensive capacity is not mentioned, the four points introduced by the US military are ambiguous with respect to their scope and limitations.

Economics and Employment

Issues Currently, 8% of jobs are occupied by robots, but in 2020 this percentage will rise to 26%. Robots will become increasingly autonomous and be able to interact, execute and make more complex decisions. Thanks to 'big data', robots now have a formidable database that allows them to experiment and learn which algorithms work best. The accelerated process of technological development now allows labor to be replaced by capital (machinery). However, there is a negative correlation between the probability of automation of a profession and its average annual salary, suggesting a possible increase in short-term inequality [Hemous& Olsen, 2016]. The problem is not the number of jobs lost through automation, but in creating enough to compensate for potential job losses. In previous industrial revolutions, new industries hired more people than those who lost their jobs in companies that closed, because they could not compete with the speed of development in new technologies [Virgillito, 2017]. An important note, about this revolution, is the fact that it is not only the manual trades likely to be automated, but also in jobs involving tasks of an intermediate nature, such secretarial, administration and other office work. To deal with this situation, it is necessary to put in place legal frameworks that make sure the benefits of automation do not solely go to the employer but are distributed equally, to guarantee the maintenance of education, health and pensions.

Human Bias in Artificial Intelligence

In an article published by Science magazine, researchers saw how machine learning technology reproduces human bias, for better or for worse. Words related to the lexical domain of flowers are associated with terms related to happiness and pleasure (freedom, love, peace, joy, paradise, etc.). The words relating to insects are, conversely, close to negative terms (death, hatred, ugly, illness, pain, etc.).

This reflects the links that humans have made themselves. Artificial Intelligence biases have already been highlighted in other applications. One of the most notable was probably Tay, a Microsoft AI launched in 2016, which was supposed to embody a teenager on Twitter,

able to chat with internet users and improve through conversations. However, in just a few hours, the program, learning from its exchanges with humans, began to make racist and anti-Semitic remarks, before being suspended by Microsoft (see Sidebar – Failures of Artificial Intelligence). The problem is not only at language level. When an Artificial Intelligence program became a juror in a beauty contest in September 2016, it eliminated most black candidates as the data on which it had been trained to identify "beauty" did not contain enough black skinned people.

Limitations and Opportunities of Artificial Intelligence

Artificial Intelligence has the potential to change the world but there are still many problems to overcome before its widespread applications. Furthermore, its practical use is not without failures. Recently, we have seen a surge of interest in deep learning with promising results that will reshape the future of AI. But deep learning is only one of the many tools that the Artificial Intelligence community has developed over the years. It is important to put into perspective the current development of Artificial Intelligence and its specific limitations.

- **Intelligence as a multi-component model:** A machine to be called "intelligent" should satisfy several criteria that include the ability of reasoning, building models, understanding the real word and anticipate what might happen next. The concept of "intelligence" is made of the following high-level components: perception, common sense, planning, analogy, language and reasoning.
- **Large datasets and hard generalization**: After extensive training on big datasets, today machines can achieve impressive results in recognizing images or translating speech. These abilities are obtained thanks to the derivation of statistical approximations on the available data. However, when the system has to deal with new situations when limited training data is available, the model often fails. We know that humans can perform recognition even with small data since we can abstract principles and rules to cover a diverse range of situations. The current Artificial Intelligence systems are still missing this level of abstraction and generalisability.
- **Black box and a lack of interpretation:** Another issue with the current Artificial Intelligence system is the lack of interpretation. For example, deep neural networks have millions of parameters and to understand why the network provides good or bad

results becomes impossible. Despite some recent work on visualising high-level features by using the weight filters in a convolution neural network, the obtained trained models are often not interpretable. Consequently, most researchers use current AI approaches as a black box.

- **Robustness of Artificial Intelligence:** Most current Artificial Intelligence systems can be easily fooled, which is a problem that affects almost all machine learning techniques.

Despite these issues, it is certain that Artificial Intelligence will play a major role in our future life. As the availability of information around us grows, humans will rely more and more on Artificial Intelligence systems to live, to work and to entertain. Therefore, it is not surprising that large tech firms are investing heavily on Artificial Intelligence related technologies. In many application areas, Artificial Intelligence systems are needed to handle data with increasing complexities. Given increased accuracy and sophistication of Artificial Intelligence systems, they will be used in more and more sectors including finance, pharmaceuticals, energy, manufacturing, education, transport and public services. In some of these areas they can replace costly human labour and create new potential applications and work along with/for humans to achieve better service standards.

It has been predicted that the next stage of Artificial Intelligence is the era of augmented intelligence. Ubiquitous sensing systems and wearable technologies are driving towards intelligent embedded systems that will form a natural extension of human beings and our physical abilities.

Human sensing, information retrieval and physical abilities are limited in a way that Artificial Intelligence systems are not. Artificial Intelligence algorithms along with advanced sensing systems could monitor the world around us and understand our intention, thus facilitating seamless interaction with each other.

Advances in Artificial Intelligence will also play a critical role in imitating the human brain function. Advances in sensing and computation hardware will allow to link brain function with human behaviour at a level that Artificial Intelligence self-awareness and emotions could be simulated and observed in a more pragmatic way. Recently, quantum computing has also attracted a new wave of interest from both academic institutions and

technological firms such as Google, IBM and Microsoft. Although the field is at its infancy and there are major barriers to overcome, the computational power it promises, potentially relevant to the field of Artificial Intelligence, is well beyond our imagination.

Conclusion and Recommendations

There are many lessons that can be learnt from the past successes and failures of Artificial Intelligence. To sustain the progress of Artificial Intelligence, a rational and harmonic interaction is required between application specific projects and visionary research ideas. Along with the unprecedented enthusiasm of Artificial Intelligence, there are also fears about the impact of the technology on our society. A clear strategy is required to consider the associated ethical and legal challenges to ensure that the society as a whole will benefit from the evolution of Artificial Intelligence and its potential adverse effects are mitigated from early on. Such fears should not hinder the progress of Artificial Intelligence but motivate the development of a systematic framework on which future AI will flourish. Most critical of all, it is important to understand science fiction from practical reality. With sustained funding and responsible investment, Artificial Intelligence is set to transform the future of our society - our life, our living environment and our economy.

The following recommendations are relevant to the research community, industry, government agencies and policy makers:

- Robotics and Artificial Intelligence are playing an increasingly important role in the UK's economy and its future growth. We need to be open and fully prepared for the changes that they bring to our society and their impact on the workforce structure and a shift in the skills base. Stronger national level engagement is essential to ensure the general public has a clear and factual view of the current and future development of robotics and Artificial Intelligence.
- A strong research and development base for robotics and AI is fundamental to the UK, particularly in areas in which we already have a critical mass and international lead. Sustained investment in robotics and Artificial Intelligence would ensure the future growth of the UK research base and funding needs to support key Clusters/Centres of Excellence that are internationally leading and weighted towards projects with greater social-economic benefit.

- It is important to address legal, regulatory and ethical issues for practical deployment and responsible innovation of robotics and Artificial Intelligence; greater effort needs to be invested on assessing the economic impact and understanding how to maximise the benefits of these technologies while mitigating adverse effects.
- The government needs to tangibly support the workforce by adjusting their skills and business in creating opportunities based on new technologies. Training in digital skills and re-educating the existing workforce is essential to maintain the competitiveness of the developing countries.
- The UK has a strong track record in many areas of RAS and AI. Sustained investment in robotics and AI is critical to ensure the future growth of the UK research base and its international lead. It is also critical to invest in and develop the younger generation to be robotics and AI savvy with a strong STEM foundation by making effective use of new technical skills.

REFERENCES

1. Young, A. and Yung, M. (1997). "Deniable password snatching: On the possibility of evasive electronic espionage," in Security and Privacy. Proceedings, IEEE Symposium on, pp. 224-235.

2. Bryant, C. and Waters, R. (2015). "Worker at Volkswagen plant killed in robot accident," in Financial Times, ed.

3. Reilly, C. O. and Munakata, Y. (2000). Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain: MIT press.

4. Pradhan, C. (2001). Recent Developments in Analytic Philosophy, Indian Council of Philosophical Research, New Delhi, pp. 338-340.

5. Rogers, C. (2015). "Google Sees Self-Driving Cars on Road within Five Years," Wall Street Journal.

6. Smarr, C. A., Prakash, A., Beer, J. M., Mitzner, T. L., Kemp, C. C., and Rogers, W. A. (2012) "Older adults' preferences for and acceptance of robot assistance for everyday living tasks," in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, pp. 153-157.

7. Floreano, D. and Wood, R. J., (2015). "Science, technology and the future of small autonomous drones," Nature, vol. 521, pp. 460-466.

8. Hémous, D. and Olsen, M. (2016) "The Rise of the Machines: Automation, Horizontal Innovation and Income Inequality".

9. Kirat, D., Vigna, G. and Kruegel, C. (2014). "Barecloud: bare-metal analysis-based evasive malware detection," in 23rd USENIX Security Symposium (USENIX Security 14), , pp. 287- 301.

10. Broadbent, E., Stafford, R. and MacDonald, B. (2009). "Acceptance of healthcare robots for the older population: review and future directions," International Journal of Social Robotics, vol. 1, pp. 319-330.

11. Horvitz, E. (2014)."One Hundred Year Study on Artificial Intelligence: Reflections and Framing," ed: Stanford University.

12. Hatfield, G. (2003). Descartes and the Meditations, Rutledge, New York. p. 120.

13. Ryle, G. (1985). The Concept of Mind, Penguin Books Ltd., Middlesex, England, p. 17.

14. Arisumi, H., Miossec, S., Chardonnet, J. S. and Yokoi, F. (2008). "Dynamic lifting by whole body motion of humanoid robots," in Intelligent Robots and Systems, IROS, IEEE/RSJ International Conference on, pp. 668-675.

15. Lighthill, I. (1973). "Artificial Intelligence: A General Survey," in Artificial Intelligence: A Paper Symposium. London: Science Research Council.

16. Andreu-Perez, J., D., Leff, R. F., Shetty, K., Darzi, A. and Yang, G. Z. (2016). "Disparity in Frontal Lobe Connectivity on a Complex Bimanual Motor Task Aids in Classification of Operator Skill Level," Brain connectivity, vol. 6, pp. 375-388, 2016.

17. Gonzalez, J. and Barr, V. (2000). "Validation and verification of intelligent systems-what are they and how are they different?," Journal of Experimental & Theoretical Artificial Intelligence, vol. 12, pp. 407-420.

18. Harrison, J., Izzetoglu, K., Ayaz, H., Willems, B., Hah, S., Ahlstrom, U. et al., (2014). "Cognitive workload and learning assessment during the implementation of a next-generation air traffic control technology using functional near-infrared spectroscopy," IEEE Transactions on Human-Machine Systems, vol. 44, pp. 429-440.

19. Haugeland, J. (1989). Artificial Intelligence: The Very Idea, MIT Press, USA, p. 2.

20. Searle, S. R. (1987). “Minds and Brains without Programs,” in Mindwaves: Thoughts on Intelligence, Identity and Consciousness, C. Blakemore, and S. Greenfield (ed.), Basil Blackwell, Oxford, p. 210.

21. Russell, J. and Norvig, P. (2009). Artificial intelligence: a modern approach (3rd edition): Prentice Hall.

22. William, J. (1948). Essays in Pragmatism, ed. A. Castell (Hafner, New York), P. 73

23. Mochizuki, K., Nishide, S., Okuno, H. G. and Ogata, T. (2013). "Developmental human-robot imitation learning of drawing with a neuro dynamical system," in Systems, Man, and Cybernetics (SMC), IEEE International Conference, pp. 2336-2341.

24. Zhang, L., Jiang, M., Farid, D. and Hossain, M. (2013). "Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot," Expert Systems with Applications, vol. 40, pp. 5160-5168.

25. Asada, M. (2015). "Towards artificial empathy," International Journal of Social Robotics, vol. 7, pp. 19-33.

26. Virgillito, M. E. (2017). "Rise of the robots: technology and the threat of a jobless future," Labor History, vol. 58, pp. 240-242.

27. Pamela, M. (1979). This view quoted by, in his Machines Who Thinks, W.H. Freeman and Company, San Francisco, p. 70.

28. Chan, M.T., Gorbet, R., Beesley, P. and Kulic, D. (2015)"CuriosityBased Learning Algorithm for Distributed Interactive Sculptural Systems," in Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference, pp. 3435-3441

29. Mavridis, N. (2015). "A review of verbal and non-verbal human–robot interactive communication," Robotics and Autonomous Systems, vol. 63, pp. 22-35.

30. Spinrad, P. (2017). "Mr Singularity," Nature, vol. 543, pp. 582-582.

31. Alston, P. (2011). "Lethal robotic technologies: the implications for human rights and international humanitarian law," JL Inf. & Sci., vol. 21, p. 35.

32. Oudeyer, P. Y. (2014). "Socially guided intrinsic motivation for robot learning of motor skills," Autonomous Robots, vol. 36, pp. 273-294.

33. Descartes, R. (1984). The Philosophical Writing of Descartes, Vol. II, Edited and Translated by John Cottingham, Robert Stoothoff, Dougald Murdoch, p. 13.

34. Descartes, R. (1984). The Philosophical Writing of Descartes, Vol. II, Edited and Translated by John Cottingham, Robert Stoothoff, Dougald Murdoch, Cambridge University Press, Cambridge, p. 19.

35. Descartes, R. (1984). The Philosophical Writing of Descartes, Vol. II, John Cottingham, Robert Stoothoff, Dougald Murdoch, p. 54.

36. Descartes, R. (1984). The Philosophical Writing of Descartes, Vol. II, Edited and Translated by John Cottingham, Robert Stoothoff, Dougald Murdoch, 1984, p. 7.

37. Descartes, R. (1985). The Philosophical Writing of Descartes, Vol.I, Edited and Translated by John Cottingham, Robert Stoothoff, Dougald Murdoch, p. 304.

38. Descartes, R. (1985). The Philosophical Writing of Descartes, Vol. I, Edited and Translated by John Cottingham, Robert Stoothoff, Dougald Murdoch, p. 140.

39. Descartes, R. (2003). Discourse on Methods, Part-V, Clarke, Desmond M. (trans), Descartes on Method and Related Writings, Penguin Books Ltd., England, pp. 24-30.

40. Lohr, (2012). "The age of big data," New York Times, vol. 11.

41. Ratschan, S. and She, Z. (2005)."Safety verification of hybrid systems by constraint propagation based abstraction refinement," in International Workshop on Hybrid Systems: Computation and Control, pp. 573-589.

42. T. E. P. s. L. A. Committee, (2016). "European Civil Law Rules in Robotics,".

43. Hobbes, T. (2003). “Elements of Philosophy,” quoted in Lilli Alanen, Descartes’ Concept of Mind, Harvard University Press, Cambridge, Mass., p. 98.

44. Kruse, T., Pandey, A. K., Alami, R. and Kirsch, A. (2013). "Human aware robot navigation: A survey," Robotics and Autonomous Systems, vol. 61, pp. 1726-1743,. ˇ Artificial Intelligence and Robotics // 46

45. Ohmura, Y. and Kuniyoshi, Y. (2007). "Humanoid robot which can lift a 30kg box by whole body contact and tactile feedback," in Intelligent Robots and Systems, IEEE/RSJ International Conference on, 2007, pp. 1136-1141.

46. Chen, Z., Jia, X., Riedel, A. and Zhang, Z. (2014). "A bio-inspired swimming robot," in Robotics and Automation (ICRA), 2014 IEEE International Conference, pp. 2564-2564.

47. Kappassov, Z., Corrales, J.-A and Perdereau, V. (2015). "Tactile sensing in dexterous robot hands—Review," Robotics and Autonomous Systems, vol. 74, pp. 195-220.