Extended Essay

Michael Cordover
bxg674
of 001701

August 23, 2004

The effects of Gödel's incompleteness theorems on the thesis that artificial intelligence can be created

An Extended Essay in Philosophy

3874 words

Abstract

What are the effects of Gödel's incompleteness theorems on the thesis that artificial intelligence can be created? Issues arising from the consideration of artificial intelligence in light of Gödel's unprovability theorems are considered. Initially the argument put forth by J.R. Lucas that a Gödel sentence could be known to be true by a human but not a machine is addressed. This involves an examination of both Gödel sentences and the ambiguity of an ability to know the truth of the sentence by different means. The addressing of this issue leads to Searle's Chinese Room problem which is then addressed. Addressing this required in-depth consideration of understanding and the functioning of the mind, either as a material or non-material entity. The philosophical theoretical implications of these are discussed but not practical problems involved in the development of artificial intelligence. Certain assumptions about epistemology and how computers can be considered to be affected by Gödel's theorems (which only affect some axiomatic systems) are made. These topics are not considered in detail in the body of the essay, rather being justified in appendices. The effect of the Gödel's theorems on the thesis that artificial intelligence can be created is finally considered to be negligible, although significant other problems are faced by the artificial intelligence development community.

Contents

1 Introduction and definitions

Artificial intelligence is a possibility that has greatly intrigued mankind. The assertion that some machine — built by a human being — could emulate perfectly the life of a human has been under great contention for many years amongst the greatest philosophers. There seemed to be some mathematical backing for the negative of that topic by invoking Gödel's incompleteness theorems to demonstrate how machines are provably unable to carry out certain tasks which humans can. There remains though some contention as to whether or not this is indeed the case. This essay will attempt to discern the effects of Gödel's incompleteness theorems on the thesis that artificial intelligence can be created.

1.1 Intelligence

Before embarking upon an examination of these arguments some definitions are required. Artificial intelligence occurs where a machine emulates a human being in pursuits of the mind. This is not only in pure reason, but also in terms of emotion and communication. This is similar to the Turing test [Turing 1950] for intelligence. It proposes, in essence, that if a human communicating with some entity can be convinced that the entity is human when it is in fact a machine then that machine will have passed the Turing test and can be considered intelligent. While in this essay no specific test of artificial intelligence is proposed the aim remains to emulate human mental behavior as accurately as possible.

1.2 Machines

A further difficulty exists about what is allowable as a machine. Many materialists argue that the human brain, centre of all mental activity, is mechanical. This extreme definition places humans as "artificially intelligent" beings. Again, to restrict, [Turing 1950] presents a reasonable solution. A Turing machine is any machine which can perform one or more recursive or otherwise well-formed algorithmic operation. In essence, a Turing machine is a single-function computer, able to complete only one task. A modern computer is closer to a Universal Turing machine, i.e. one that can emulate the function of any Turing machine when correctly programmed. For this essay a machine, the word used interchangeably with 'computer,' shall be a Universal Turing machine. For a justification of the consideration of computers as Universal Turing machines, see Appendix A.

1.3 Gödel

The more astute readers will be curious at this point as to where the incompleteness theorems fit into this essay. The Gödel incompleteness theorems were developed by Gödel in the 1930s and show the fundamental weaknesses in axiomatic systems [Gödel 1934]. In any system where the natural numbers can be defined as a set, Gödel showed that there will be an undecidable sentence which is meaningful but paradoxical. The so called Gödel sentences are essentially formal demonstrations of the liar paradox, "this sentence is false." The importance of Gödel's discovery was that it worked for the vast majority of arithmetic systems. This includes modern algebra. Gödel showed that these finite axiomatic systems were all incomplete — there would always be a contradiction in the form of and unprovable yet meaningful sentence.

Another impact of Gödel's theorems was that there could be no solution to the Gödel sentences within a related system, i.e. it is impossible to bypass this difficulty by connecting several systems and circularly solving the indeterminate sentences. Given that a Universal Turing machine can perform any task which can be performed by a Turing machine, this effectively means it can perform a large set of operations on a specific alphabet[1]. This means that for the purpose of specifying those things which are possible — and impossible — tasks for the machine one can equate the machine to an axiomatic system. Equivalently, all machines (recalling the specific definition given above) cannot prove a Gödel sentence.

[1]The word 'alphabet' is used here in the formal logical sense of a set of symbols manipulated using logic within a finite axiomatic system

1.4 Coherentism

Throughout this essay the coherentism is considered to be the way of looking at human knowledge. This is based on selection from two options - coherentism, where each piece of knowledge is intertwined and dependent upon each other, or foundationalism, where there are a few pieces of absolutely certain knowledge upon which all other pieces of knowledge are based. It seems more natural that humans would compare newfound knowledge to all of their current knowledge before accepting it, as opposed to simply trying it against a few axioms. These axioms are certainly flexible person-to-person as there is little consensus on what any foundation would be if one existed.

Logic is purely a means of preserving truth - from pure logic nothing truly new can arise. If humans do have foundational knowledge base it appears as less declarative and more procedural — rather than there being specifically axiomatic it specifies a means of comparison which is almost identical for most people. To put it into programming terms, the same function set with different data. The question of foundationalism or coherentism for knowledge though is a long one, barely touched here. More information is available in Appendix B.

2 Lucas' conjecture

2.1 Description

Gödel sentences are often given as "this sentence is unprovable within the system." A crude, informal wording but effective nonetheless. As was initially suggested by [Lucas 1961], this is quite clearly true. Any human with an understanding of the language can understand clearly that the sentence is unprovable — thus it is true. Given that machines operate wholly as (Universal) Turing machines they cannot show, nor know[2], that the sentence is true. Here Lucas believed he'd found an example where a machine could not perform a task a human could with relative ease. This was claimed as

[2] Forgive the anthropomorphism, but this *is* an essay on artificial intelligence

a demonstration of a fundamental theorem of computing which prevented artificial intelligence.

To react though, consider a series of systems, each subtly different from the last, each able to solve the Gödel sentence of another. Certainly, the theorem shows that some contradiction will exist — but it is a common human trait to hold conflicting ideas simultaneously. This, however, is not how it works. The further contradiction that arises will be another Gödel sentence — trivially easy for humans and impossible for purely logical entities, i.e. machines.

2.2 Human-computer similarity response

Human knowledge exists as an interdependent web, with each fact being verified because it fits in with everything else known. This seems to be in stark contrast to the system used by machines, which appears fundamentally axiomatic. Yet upon close examination the system used by computers is one of recall and check — the axioms are not fixed, for a slightly (or even dramatically) different system can be used for each operation.

Of course it would be trivial to train a computer to recognise statements in a Gödel sentence form and reject them, or call them true, but this is arguably not the computer thinking. This is too much human intervention. Yet all humans are trained in things, from birth, shown what is right and what is wrong. A machine, able to recognise and automatically answer Gödel sentences in a specific way could be considered to simply be following its training, as humans do. Many skills and behaviors humans have are simply matters of recognition and trained, algorithmic (similar for several situations) action.

The way training is used demonstrates the similarity between human knowledge manipulation and that of machines. Humans have particular means of comparison which stays constant, by which new knowledge is verified. If something is perceived but does not fit into the existing knowledge of a person it will generally be rejected. Machines, similarly, use a certain set of deductive logical laws in order to ascertain the validity of knowledge relative to other knowledge. This is a similarity to human operation.

The objection of course arises that this is conditioning by the programmer of the computer. This is true, yet a human is also conditioned to see the truth in a Gödel sentence. A human also faces the limitation of being able to prove a Gödel sentence — it is provably impossible to do so. A human can know it to be true, however, based on training or non-deductive reasoning, as could a machine using these methods.

2.3 Symbol manipulation response

The difficulty of course arises when attempting to allow a machine non-deductive reasoning, or some method which is not purely logical. A machine cannot do this — all a machine can do is manipulate symbols according to instructions. It is the human interpretation of these symbols which gives them meaning. A machine can be programmed to generate entire essays[3] which follow English syntax perfectly. It is even possible to develop applications which can communicate meaningfully to humans on particular topics, even if the machines cannot respond to human input. This generation of text does not use pure logic, but iterative manipulation of symbols created by humans, placed in the machine's memory by humans, manipulated as segments of memory by a purely logical machine.

3 The Chinese Room problem

3.1 Description and relevance

In [Searle 1980] the Chinese Room problem creates a situation in which the author, John, who understands English, knows no Chinese. Indeed, he is so ignorant of the language, he could not confidently tell it apart from Japanese or Thai. He is locked in a room and provided with two sheets of paper, the first written in English and the second in a language presumed to be Chinese. The English sheet provides instructions on how to create a

[3] For example, the Postmodernism generator at http://www.elsewhere.org/cgi-bin/postmodern/

third sheet, written in Chinese, based on what is written on the Chinese sheet. It is the simple task of recognising certain shapes on one sheet and then placing a new shape on the blank sheet. Once John has completed this task the third sheet is collected from him. To an observer outside the room, John is a black box which when provided with input (the second sheet) and instructions, call them a program, (the first sheet) produces output (the third sheet). Searle argues that even if John became so proficient at this process that to an outsider providing an input of questions, the output appears to be answers given by a person fluent in Chinese, John still does not understand the language. Searle claims that this is analogous to a machine and thus tries to show that machines can never be intelligent, for they cannot understand.

In the Chinese Room example John is acting purely as a symbol manipulator, as a machine can do, yet apparently has no understanding of what is being manipulated. If one were to accept the counterarguments to Lucas' conjecture presented above then this issue of understanding would still exist. The Chinese Room problem is a major objection to the rebuttals of Lucas presented above.

3.2 Understanding

The analogy presented in [Searle 1980] is flawed from the outset. While setting up the conditions, Searle says "[t]he rules are in English, and I understand these rules as well as any other native speaker of English." The key word here is "understand" — Searle admits that to carry out any instruction the machine must *understand* those instructions. This is contrary to the aim of the article - to demonstrate that machines can not have understanding. In and of itself this does not destroy Searle's argument, but it does cause the analogy to be compatible with an argument presented below, contradicting Searle.

3.3 Humans as symbol manipulators

The thesis here presented is that humans are themselves nothing more than symbol manipulators. Consider the mind to be a machine and the body to be its means of interfacing with others i.e. its input and output. That which enters the mind through the body becomes some symbol of itself within the mind, neither the direct sense data (i.e. not the light which enters the eye) nor the object itself. It must, therefore, be some representation. This representation is used purely internally (evidence for this could be taken from the fact that the same object might be called two different things in two different languages) and hence could be arbitrary. If the internal descriptor could be arbitrary, the mind is acting as John inside the body. The input is in the form of sensory data, the output in several forms but generally as some responsive action.

Rather than refuting the claim of Searle's that machines cannot have understanding, here is presented a refutation of the idea that humans have understanding as Searle claims it. Thus, machines are capable of the same amount of understanding as humans, so can be artificially intelligent.

There are significant objections to this claim. Understanding is, after all, defined by what humans can do. The idea that we have no understanding of anything seems ludicrous. Yet consider what this really proposes: there is an internal language of thought, which cannot be articulated in any human language. Each concept known to each person has an arbitrary "word" which is used to refer to it within the mind. These "words" might of course be no more than a particular sequence of neural firings, or they might be something non-material, depending on the solution to the mind-body problem used (materialism, idealism or something in between). Regardless of what these words are it remains visible that they exist and further that what these representations are can be arbitrary. They are manipulated within the mind, a system which is internal.

For the purpose of example, consider that each concept is given a sequential number based on when it was learned. For simplicity, consider nouns. The concept of "truck" might have the reference 1, "wheel"→2 etc.

Any thoughts about trucks would use 1, actions within the mind would be performed on the number 1 rather than on the word "truck." A vocabulary could be built up where 3 is the concept of having something — e.g. 3(1,2) means 'truck has wheel.' The concept labeled 0 might be the imperative of saying, $10x$ be the word for a concept x. From this conditional sentences could be constructed, e.g., `IF 3(1,2) THEN { 0(101);0(103);0(102); }` which would say cause the body to say "truck has wheel" if a truck has a wheel. This example is exceedingly simple but demonstrates the idea of reference.

Where though, does the structure come from? The `IF/THEN` conditional, the idea of verbs accepting nouns, etc. are all implicit in the definition above. They must originate somewhere. It cannot be through conditioning or training, since all knowledge is being justified by that system. In a modern digital computer this arises from the use of the transistor switch, a small piece of hardware which can perform an `IF/THEN` operation. From this all other operations can be developed and it is here proposed that the mind has a similar system. Synapses within the brain indeed operate in a fashion similar to transistors — with a large enough current, the signal flows, but if there is not a large enough current the signal does not flow. Systems with the brain are such that the number of discrete flows within a given time determine the strength of the signal, rather than there being an analogue signal strength directly used [Beckett 1986]. This is very similar to the systems used in digital electronics to emulate uncertainty and other analogue quantities.

Evidence for the claim of a separate language of thought has been provided, however, only under a materialist account of the mind-body problem, i.e. that the mind is the purely physical brain. Such strong materialist accounts are reasonably rare. There is some — albeit weaker — evidence that a non-material mind would operate in a similar manner: different people may have different associations for a particular object, implying that they assigned relationships to that object's internal pointer differently. This does not, however, offer evidence that the mind is able to perform conditional operations natively, although empirically this seems the case. Examples of people who do not have the ability to formulate or understand conditionality are rare if extant at all. It would seem that an understanding of the conditional, i.e. the `IF/THEN` idea, is required for intelligence.

The two premises thus presented to demonstrate that humans are merely symbol manipulators are: firstly, that there is an internal and arbitrary language of thought; secondly, that conditional manipulation exists within the mind without training — i.e. the `IF/THEN` ability of humans is innate from birth.

3.4 Qualia

The major objection to the idea that humans do nothing more than manipulate symbols is the idea of qualia. Qualia is the experience of thought itself, perhaps easier known as an awareness of thought. This is a sort of meta-thinking (i.e. thinking about thinking) which occurs without active thought. The language to describe qualia is difficult because of its self-referential nature. It refers to the knowing that there is some thinking going on, rather than the active thinking about the act of thinking generally. It is this claimed ability to know the state of the mind which shows that there is something more than the menial manipulation of concepts going in the mind.

Empirically, however, there is a contradictory argument. Each time a state of mind is described it is described in its past sense — the passage of time is such that there is no way to consider, discuss or think about a state of mind in the present. Attempting to think about the state of mind one has would alter that state of mind, hence what is thought about is a concept of a particular state of mind. This concept, similar to a memory, is a the collection of things which were being thought about at the time. This is equivalent to thinking about the same thing while adding other influences to it. Alternately, one might consider that it is not a collection of things which were being thought about, but an association with the actions which could be taken based on that thought.

The important section of this idea remains that all thinking about thinking is done on a concept of thought itself, which could hold its own symbol under the model presented in §3.3. This symbol can then be "thought about," i.e. used in conjunction with conditionals and other symbols to develop an outcome. The difficulty experienced in understanding and expressing qualia would be then justified as it is a concept which exists purely internally, without reference to the outside world. While it may make sense within the internally consistent[4] system of the mind it cannot be related to the tangible world, so is difficult to express in terms of those tangible-world concepts.

Proponents of qualia will continue to claim that there is some awareness of thought while being unable to specify it other than in back reference. Thinking about a specific thought is necessarily about a past thought, which can have been assigned a specific "word." It is not possible to know about the current state of the mind, for to do so would be to modify the state of the mind. This is, perhaps ironically, the self-referential problem which is exploited in most paradoxes, including the one proposed by Gödel. To know directly the current state of mind is impossible, but to think about a past state of mind, or states of mind in general, is simple and requires only a knowledge of the vocabulary which is used to refer to each concept that can be thought about.

4 Concluding remarks

Gödel's unprovability theorems show that few axiomatic systems are complete and consistent — indeed that arithmetic is an inconsistent system. J.R. Lucas, drawing on the fact that machines extensively use arithmetic systems, developed an argument in his paper *Minds, Machines and Gödel* that there could be no artificial intelligence. He used the structure of the Gödel sentence, which is obviously true to people, yet not provable in the axiomatic system in which it occurs, to demonstrate a capability of humans which computers could never have: to know that a Gödel sentence was true.

Lucas' proposition failed to acknowledge that neither a human nor a

[4]Not necessarily completely consistent, but as consistent as the system of knowledge will allow

machine could prove a Gödel sentence to be true within the system in which it occurs. Both a machine and a human, however, can be trained to recognise the truth in the sentence. This is because while a machine can only manipulate within an axiomatic system is can manipulate symbols which represent things which have meaning to humans. Thus a machine could manipulate symbols in order to appear to be intelligent while the interpretation by humans is essential to this.

The argument presented about symbol manipulation is Searle's Chinese Room problem. This uses an analogy to demonstrate the way a machine which, to an outsider observer, can seamlessly respond to human interaction still has no understanding of the concepts involved. The immediate irony is that in his explication of the analogy Searle uses the word "understand" to describe the language the 'machine' can read. This in itself is not enough to declare the analogy to be void but does create a vulnerability in the analogy.

The way that understanding is described by Searle seems to be coherent when considered to be nothing more than a specific type of symbol manipulation. This would make humans mere machines and remove the key theoretical barriers to artificial intelligence. The major counterargument to this, that qualia exists, is difficult to express properly. This difficulty of expression of precisely what qualia is and how it is to be understood creates an opening whereby qualia can be considered in the model of humans as symbol manipulators.

From the questions posed above it would seem that many of the theoretical difficulties in creating artificial intelligence which arise as a result of consideration of Gödel's unprovability theorems can be solved. Lucas' conjecture on the human provability of Gödel sentences fails to realise the restrictions on the theorems — they are applicable in a very limited set of circumstances. Even accepting that there is some provability by humans, ignoring the limitations imposed on the theorems by themselves, Lucas' argument faces severe difficulties. From it, though, the Chinese Room problem is encountered. This is one of the more resilient and often used analogies to combat artificial intelligence. Despite this it contains several fatal flaws, not the least of which is the author's own misuse of terminology for what he was trying to prove.

This essay attempts address some of the problems for artificial intelligence which arise as a result of considering artificial intelligence in light of Gödel's unprovability theorems. Some parts of this essay may not be accepted by some, perhaps other problems will be raised by others. It can be considered, however, the the effect of Gödel's unprovability theorems on the thesis that artificial intelligence can be created is negligible. Regardless the ideas presented here are all purely theoretical. There are huge technical barriers to be overcome before artificial intelligence can be created. Until it is, if ever,created opponents will present theoretical arguments rejected by proponents of artificial intelligence. It shall indubitably remain a topic of great contention for philosophers.

A Computers as axiomatic systems

There is possibly some requirement for justification of the description of a computer as several Turing machines combined — i.e. a Universal Turing machine. A Turing machine, as has been stated, is a theoretical machine which can represent, in its various states, a certain part of an axiomatic system. It can perform only a single function repeatedly upon its alphabet. A computer, alternately, is programmable to perform a theoretically infinite number of functions. Any system which can emulate a Universal Turing machine — i.e. can compute for any meaningful sentence within its alphabet — is called Turing-complete. Most programming languages for modern computers are provably Turing-complete, showing how the computer can act as a Universal Turing machine. Furthermore a Universal Turing machine can be described as a combination of Turing machines. A Universal Turing machine has the ability to emulate any Turing machine, hence any single operation carried out by a Universal Turing machine could be carried out by a single Turing machine.

A common objection arises in the form of the ability of computers to manipulate languages - rudimentary translation, document writing etc. programs exist. This would suggest that there is some further ability of computers to render language, something beyond mathematics. It is a mistake, however, to presume that the computer can understand the language it manipulates — instead it manipulates the symbols it is given as binary digits (bits) to create an output stream. This output stream, always the same for a given input stream, is then interpreted by humans to have meaning. A human interpretation does not imply, however, that the machine can understand the English (or other language, for example Chinese) that it has been given.

Some evidence that Gödel's theorems have relevance to computers may be required. This can be obtained by examining the fundamental workings of a computer. At its core a computer operates as a pure mathematical device, adding and subtracting numbers in binary form. The arrangement of these numbers is then interpreted by humans to have a specific meaning. The machine can appear to perform certain tasks which would require human thought but in fact all a modern computer can do is iterate quickly.

B Theories of knowledge

There are two major theories of knowledge — coherentism and the foundationalism. One major proponent of the former is Donald Davidson and the latter is discussed in detail in [Descartes 1901]. Coherentism is also known as "web of knowledge" theory. Foundationalism claims that there are a few basic axiomatic, self-evident facts upon which all knowledge can be based. Descartes, for example, suggested that the foundation was self-awareness in his famous phrase "cogito ergo sum" (I think therefore I am). Coherentism suggests instead that all knowledge depends on all other knowledge — justification arises from compatibility, or coherence, with an existing system of beliefs. Things which are not coherent are rejected.

Coherentism offers the distinct advantage that it models human behavior far better than foundationalism. People hold conflicting beliefs within themselves possibly because they only compare pieces of knowledge to other pieces of knowledge similar to it, as opposed to verification based only on a few axioms. Axiomatic verification would prevent this contradictory points of view being held by a single person. Another place where foundationalism falls is in the foundations themselves. There is always great contention for rational foundationalist claims that all knowledge can be deductively formed from a few simple statements. For one thing, logic is only a means of preserving truth — all statements are eventually tautological when expressed in relation to the axioms. As such it would be greatly difficult to have a small set of axioms which encompassed all knowledge.

Coherentism is not without its own flaws. The largest is that of self-reference being an impossibility of justification. Any system which is internally consistent could be valid under coherentism, which poses obvious difficulties when that system is not coherent according to others. These problems are similar to those raised by relativism.

Arguments aside, foundationalism is commonly considered to be a system in which it is easier to create artificial intelligence — once the axioms are found they can be placed in a machine and all of knowledge (especially with rational foundationalism) can be derived from that. The belief more commonly held by opponents of artificial intelligence is coherentism or something similar. Coherentism has been used in this essay for many reasons, including that it is compatible with the ideas presented while also being supported by many opponents of those ideas.

References

[Beckett 1986] Beckett, B.S., 1986: "Co-ordination in Animals" in *Biology: A modern introduction (GCSE edition)*, 1986, Oxford University Press, Oxford, p. 139-150. ISBN: 0 19 914260 2.

[Copeland 2002] Copeland, J.B., 2002: "The Church-Turing Thesis" in *The Stanford Encyclopædia of Philosophy (Fall 2002 Edition)*, ed. Zalta, E.N., URL: http://plato.stanford.edu/archives/fall2002/entries/church-turing/, accessed 22 December 2003

[Descartes 1901] Descartes, R., 1901: *Meditations* URL: http://philos.wright.edu/DesCartes/MedE.html, accessed 29 February 2004, first published 1641, France in Latin, this translation 1901 by J. Veitch with editors Taylor, C.S. and Manley, D.B.

[Gödel 1934] Gödel, K., 1934: "On undecidable propositions of formal mathematics" as reprinted 1964 in Feferman, S, et al., (eds.), 1986: *Collected Works of Kurt Gödel*, Vol 1, Oxford University Press, New York, p. 338-371. ISBN: 0 195 03964 5.

[Lucas 1961] Lucas, J.R., 1961: *Minds, Machines and Gödel*, URL: http://users.ox.ac.uk/~jrlucas/mmh.html, accessed 17 December 2003, originally published in *Philosophy*, 1961, XXXVI p. 112-127.

[Searle 1980] Searle, J.R., 1980: "Minds, Brains and Programs" URL: http://members.aol.com/NeoNoetics/MindsBrainsPrograms.html, accessed 29 February 2004, first published "Minds, Brains and Programs" *Behavioral and brain sciences* 3, p. 417-424.

[Shjapiro 1987] Shapiro, S.C.; Ecroth, D. (eds.), 1987: *Encyclopædia of Artificial Intelligence*, Vol 2, John Wiley & Sons, New York. ISBN: 0 471 80748 6.

[Turing 1950] Turing, A., 1950: "Computing Machinery and Intelligence," *Mind* 59 p. 433-460

[Zalta 2003] Zalta, E.N., 2003: "Turing Machine" in *The Stanford Encyclopædia of Philosophy (Summer 2003 Edition)*, ed. Zalta, E.N., URL: http://plato.stanford.edu/archives/sum2003/entries/turing-machine/, accessed 26 February 2004.