The Unexpected Hanging Paradox from an AI Viewpoint

Matjaž Gams
Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
E-mail: Matjaz.Gams@ijs.si, dis.ijs.si/mezi

Position paper

Keywords: logical paradox, artificial intelligence, progress of human civilization

Received: June 6, 2014

This position paper hypothesizes that humans are becoming smarter, not only when using ICT and AI tools, but on their own, particularly due to the progress of AI knowledge. As is common when demonstrating that one computing mechanism is stronger than another, we chose a well-known task – the unexpected hanging paradox –that humans were previously unable to resolve efficiently, but can now do so thanks to new knowledge. We show that the cause of prior problems was with ambiguous definition, as it was in the case of the liar paradox.

Povzetek: Predstavljena je hipoteza, da ljudje postajamo čedalje pametnejši zaradi spoznanj umetne inteligence, pokazana na paradoksu nepričakovanega obešanja.

1 Introduction

According to the Flynn effect [1], scores on the standard broad-spectrum IQ tests improve by up to three IQ points each decade, and the gains are even higher in some specialized areas. One theory claims that the increase of human intelligence is related to the use of information tools [2], which often progress exponentially over time.[3]

This paper presents a tentative hypothesis that artificial intelligence (AI) influences human intelligence in a positive way; specifically, it increases the ability to solve mental problems. We illustrate the hypothesis in Figure 1. The *y* axis is logarithmic in the scale. Therefore, the linear growth of computer skills on the graph corresponds to the exponential nature of Moore's law.[4] Basic human physical and mental properties, such as speed of movement, coordination or speed of human computing, have remained nearly constant in recent decades, as represented by the horizontal line in Figure 1.

Our first thesis is that, analogous to mechanical machines that enable humans to move faster than on their own, the ability of humans to solve problems increases due to information tools such as computers, mobile devices with advanced software, and AI in particular (the bold top line in the Figure 1). (The overall human ability to solve problems is growing, due to a number of reasons, primarily the growth of ICT capabilities, or advances in computers, mobile devices, and the Web.) Programs such as the Google browser may provide the greatest knowledge source available to humans, thereby representing an extension of our brains.

We go a step further in this paper. Whereas mechanical machines do not increase our physical capabilities, human intelligence generally increases on its own. For example, not only does a person play better chess when using advice from online chess programs, they also perform better when playing against other human opponents. This is due to previous interactions with chess-playing programs. In the AI community [5], it is generally accepted that AI progress is increasing and might even enable human civilization to take a quantitative leap.[6]

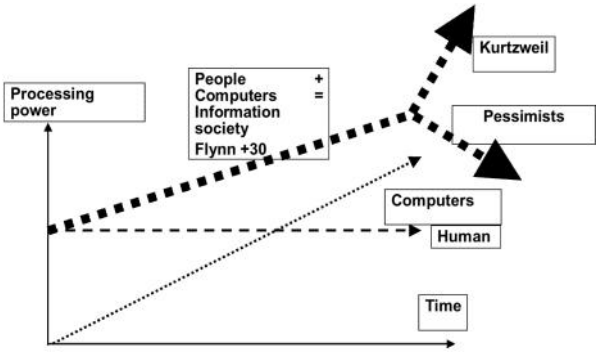

Figure 1: Growth of computer and human computing powers.

Several opposing theories claim that humans actually perform worse on their own, since machines and tools have replaced humans' need to think on their own. We argue that while this effect may be valid for human physical properties such as obesity, it is not the case in mental tasks. Another pessimistic viewpoint suggests that intelligent civilizations decline after reaching a certain development level (see Figure 1), possibly because of overpopulation, self-destruction or depletion of natural resources. This would explain why we have not yet detected alien civilizations, though the Drake's equation [7] indicates that many such civilizations should exist.

In real life, ICT, and AI tools in particular, have already significantly modified the way humans exercise their mental activities. For example, professional chess players intensively use computer chess programs to analyze game strategies and improve their level of play. Furthermore, computers outperform the best humans in nearly all mental games, with some rare exceptions such as Go. Therefore, this online advice helps humans play much better than on their own. Although it is safe to claim that computers have already significantly improved human gaming performance, is the phenomenon valid in other areas?

If we can show that humans can solve logical puzzles that they were not able to solve until recently without computers, that would be a good indication of humans getting smarter on their own. One way to confirm this idea would be to analyze the logical solutions that humans solved in the last decade. Another way would be to provide a new solution to an existing puzzle. One objection might be that just one solution of one puzzle is far too little to show anything. On the other hand, since the author of this paper is a well-educated AI scientist and not a professional logician, it could provide a reasonable indication that the tentative idea might be valid.

To demonstrate the idea, we analyze the unexpected hanging paradox.[8, 9, 10, 11] In addition, we discuss if AI programs would crash from such well-known logical paradoxes or resolve them.

2 The liar paradox

First, however, we quickly investigate the liar paradox (in which a liar says that he is a liar), first published in [12]. According to [13] it was first formulated by the Greek philosopher Eubulides of Miletus: “A man says that he is lying. Is what he says true or false?” This sentence is false when it is true. It supposedly leads to a paradox and causes logical AI machines to crash, such as in the “I, Mudd” episode of the science-fiction television series *Star Trek*.

However, as Prior shows [14], there is no paradox, since the statement is false. It is a simple contradiction of the form “A and not A,” or “It is true and false.” In other words, **if a person always lies by definition, then that person is, by definition, not allowed to say anything that is not a lie**. Therefore, such statements are simply not allowed, which means they are false. In summary, no decent AI computing machine should fail to see the falsity of the liar paradox sentence.

How did the liar paradox cause such attraction? An explanation at hand is that humans fall into a loop of true/untrue derivations without observing that their thinking was already falsified by the declaration of the problem. It seems a valid logical problem, so humans apply logical reasoning. However, the declaration of the logical paradox was illogical at the start rendering logical reasoning meaningless.

In another example, 1 + 1 = 2, and we all accept this as a true sentence without any hesitation. Yet, one liter of water and one liter of sugar do not combine to form two liters of sugar water. Therefore, using common logic/arithmetic in such a task is inappropriate from the start.

The principle and paradox of multiple-knowledge [15] tentatively explain why humans easily resolve such problems. We use multiple knowledge/ways of thinking not only in parallel, but also with several mental processes interacting together during problem-solving. Different processes propose different solutions, and the best one is selected. The basic difference in multiple-knowledge viewpoint compared to the classical ones occurs already at the level of neurons. The classical analogy of a neuron is a simple computing mechanism that produces 0/1 as output. In the multiple viewpoint, each neuron outputs 2^N possible outcomes, which can be demonstrated if N outputs from a single neuron are all connected to N inputs of another neuron. In summary, the multiple-knowledge principle claims that the human computing mechanism at the level of a neuron is already much more complex than commonly described, and even more so at the level of higher mental processes.

According to the principle of multiple knowledge, humans have no problems computing that one apple and one apple are two apples, and 1 liter of water and 1 liter of sugar is 1.6 liters of liquid and a mass of 2.25 kilograms, since they use multiple thinking. A person who logically encounters the sugar-water merge for the first time may claim that it will result in 2 liters of sugar water. However, after an explanation or experiment, humans comprehend the problem and have no future problems of this kind.

Another AI solution at hand uses contexts. In arithmetic, 1 + 1 = 2. In merging liquids and solid materials, 1 + 1 ≠ 2. In the first case, the context was arithmetic and in the second case, merging liquids and solid materials. The contexts enable an important insight into the unexpected handing paradox.

3 The unexpected hanging paradox

Unlike the liar paradox, the unexpected hanging paradox (also known as the hangman paradox, the unexpected exam paradox, the surprise test paradox, or the prediction paradox) yields no consensus on its precise nature, so a final correct solution has not yet been established.[9] This is a paradox about a person’s expectations about the timing of a future event that they are told will occur at some unexpected time.[16]

The paradox has been described as follows [9]:

A judge tells a condemned prisoner that he will be hanged at noon on one weekday in the following week but that the execution will be a surprise to the prisoner. He will not know the day of the hanging until the executioner knocks on his cell door at noon that day.

Having reflected on his sentence, the prisoner draws the conclusion that he will escape from the hanging. His reasoning is in several parts. He begins by concluding that the "surprise hanging" can't be on Friday, as if he hasn't been hanged by Thursday, there is only one day

left - and so it won't be a surprise if he's hanged on Friday. Since the judge's sentence stipulated that the hanging would be a surprise to him, he concludes it cannot occur on Friday.

He then reasons that the surprise hanging cannot be on Thursday either, because Friday has already been eliminated and if he hasn't been hanged by Wednesday night, the hanging must occur on Thursday, making a Thursday hanging not a surprise either. By similar reasoning he concludes that the hanging can also not occur on Wednesday, Tuesday or Monday. Joyfully he retires to his cell confident that the hanging will not occur at all.

The next week, the executioner knocks on the prisoner's door at noon on Wednesday — which, despite all the above, was an utter surprise to him. Everything the judge said came true.

Evidently, the prisoner miscalculated, but how? Logically, the reasoning seems correct. While there have been many analyses and interpretations of the unexpected hanging paradox, there is no generally accepted solution. The paradox is interesting to study because it arouses interest in both laymen and scientists. Here, we provide a different analysis based on the viewpoint of cooperating AI agents [16][5], contexts and multiple knowledge.[15]

The prediction of hanging on one out of five possible days is well defined through a real-life empirical fact of a human life being irreversibly terminated. However, the surprise is less clearly defined. If it denotes cognitive surprise, then the prisoner can be sure that the hanging will take place on the current day. No surprise is assured each new day, even on the first day, so hanging under the given conditions is not possible. Such an interpretation makes no sense. To avoid the prisoner being cognitively certain, the following modifications are often proposed [9]:

The prisoner will be hanged next week, and the date (of the hanging) will not be deductible in advance from the assumption that the hanging will occur during the week (A).

The prisoner will be hanged next week and its date will not be deducible in advance using this statement as an axiom (B).

Logicians are able to show that statement (B) is self-contradictory, indicating that in this interpretation, the judge uttered a self-contradicting statement leading to a paradox.

Chow [10] presents a potential explanation through epistemological formulations suggesting that the unexpected hanging paradox is a more intricate version of Moore's paradox [9]:

A suitable analogy can be reached by reducing the length of the week to just one day. Then the judge's sentence becomes: "You will be hanged tomorrow, but you do not know that."

Now we can apply AI methods to analyze the paradox. First, the judge's statement is a one-sided contract (an agreement can always be written in the form of a contract) from an AI agent viewpoint, defining a way of interacting and cooperating. As with any agreement/contract, it also has some mechanisms defining the consequences if one side violates the agreement. Since the judge unilaterally proclaimed the agreement, he can even violate it without any harm to him, whereas the prisoner's violations are punished according to the judge's will and corresponding regulations. For example, if the prisoner harms a warden, the deal is probably off, and the hanging can occur at the first opportunity, regardless of whether it is a surprise. This is an introductory indication that the hanging paradox is from the real world and that it matters, and is not just logical thinking. Even more important, it enables a valid conclusion that **any error in prisoner's actions releases the judge from his promise**.

On the other hand, the judge is, by definition, an honest person and as long as the prisoner abides to the appropriate behavior, the judge will keep his word and presumably postpone the execution if the prisoner predicts the exact day of the hanging. Now, we come to the crucial definition ambiguity. The term *deducible* means that the prediction will be 100 percent guaranteed accurate about a one-time event (that is, hanging), so such **a prediction can be uttered only once a week, not each day anew**. Therefore, the prisoner has exactly one chance of not only predicting, but also **explaining with certainty to the judge**, why the hanging will occur on that particular day. The judge will have to be persuaded; that is, he will have to understand and accept the prisoner's line of reasoning. If not, the deal is off and the judge can choose any day while still keeping his word.

For further understanding of *deducible*, consider a case in which the prisoner is given a life-saving coupon on which he writes the predicted day and stores it in the judge's safe on Monday morning with the explanation attached. Obviously, the prisoner stands no chance if the judge orders handing on Monday. Namely, if the prisoner proposes Monday, he cannot provide a deducible explanation why the handing will happen on Monday. Yes, he will not be surprised in cognitive terms, but both a correct prediction and a deducible explanation are required in order to avoid hanging. The only chance to avoid hanging is to predict Friday and hope that he will not be hanged till Friday. (In this case, the judge could still object that, on Monday for example, the prisoner could not provide a plausible explanation for Friday. Yet, that would not be fair since, on Friday, the prisoner would indeed be sure of the judge coming into contradiction.) Even if the prisoner is allowed to deposit the one and only coupon on any day in the week, there is no major difference in terms of explanation in this paper. Again, if the prisoner is allowed to deposit the coupon each day anew, this formulation makes no sense.

To explain the error in the prisoner's line of reasoning (that is, logical induction), assume that instead of giving his ruling five days in advance, he gave it on Thursday morning, leaving a two-day opportunity. Since the prisoner could use the single pardon (remember: *deducible* for a one-time event means one prediction once) and save himself on Friday, he concludes that Thursday is the only day left and cashes in his only coupon with a 100 percent certain logical explanation on Thursday. However, in this case the judge could carry out the hanging on Friday. Why? Because the prisoner provided the only 100 percent certain prediction in the form of a single life-saving coupon on Thursday, which means that on Friday he could not deliver the coupon. In other words, the prisoner wrongly predicted the hanging day and therefore violated the agreement.

It turns out that the situation on Thursday is similar to the situation on Monday. Even if the judge knocks on the door on Thursday, and the prisoner correctly predicted Thursday, he still could not provide a 100 percent certain explanation why the hanging would occur on Thursday since the judge could come back on Friday as described in the above text; therefore, the judge can proceed on Thursday without violating his proclamation.

What about AI machines? Will they crash or fail as was supposed to be the case with the liar paradox? Similarly to the liar paradox, the principle of multiple knowledge provides a simple solution that AI machines should be able to compute. If both lines of reasoning (from Friday to Monday or from Monday to Friday) are simulated with some tests, the solutions should be obtained. One does not need to understand why one line of reasoning is wrong in order to operationally solve the puzzle. The AI machine can simply evaluate both of them and accept the more plausible one. However, current AI systems are not yet capable of understanding the explanation in this paper since they behave poorly on any task demanding real-life semantics.

4 Discussion

Wikipedia offers the following statement regarding the unexpected hanging paradox [9]:

There has been considerable debate between the logical school, which uses mathematical language, and the epistemological school, which employs concepts such as knowledge, belief and memory, over which formulation is correct.

According to other publications [8], this statement correctly describes the current state of scientific literature and the human mind.

To some degree, solutions similar to the one presented in this paper have already been published.[8–9] However, they have not been generally accepted and, in particular, have not been presented through AI means. Namely, AI enables the following explanation:

The error in the prisoner's line of reasoning occurs when extending his induction from Friday to Thursday, as noted earlier, but the explanation in this paper differs. The correct conclusion about Friday is not:

"Hanging on Friday is not possible" (C),

but :

"**If** not hanged till Friday **and** the single prediction with explanation was not applied for any other day before, **then** hanging on Friday is not possible." (D)

The first condition in (D) is part of common knowledge. The second condition in (D) comes from common sense about one-sided agreements: every breach of the agreement can cause termination of it. An example would be promising a one-sided reward to a person for predicting an outcome of a sporting event and then realizing that the person deposited two predictions.

The two conditions reveal why humans have a much harder time understanding the hanging paradox, compared to the liar paradox. The conditions are related to the concepts of ***time*** **and** ***deducibility*** **and should be applied simultaneously**, whereas only one insight is needed in the liar paradox. In AI, this phenomenon is well known as the **context-sensitive reasoning** (often related to agents), which was first presented in [18] and has been used extensively in recent years. Here, as in real life, under one context the same line of reasoning can lead to a different conclusion compared to the conclusion under another context (remember the sugar water). But one can also treat the conditions in statement (D) as logical conditions, in which case the context can serve for easier understanding. The same applies to the author of this paper: Although he has been familiar with the hanging paradox for decades, the solution at hand emerged only when the insight related to the contexts appeared.

Returning to the motivation for analysis of the unexpected hanging paradox, the example was intended to show that humans have mentally progressed to see the trick in the hanging paradox, similar to how people became too smart to be deceived by the liar paradox.

There are several potential objections. First, one needs no AI or ICT knowledge to see the proposed solution. However, this is the only major change of the author's knowledge from the years before the recent progress of AI knowledge. It is not only that using AI knowledge helped solve the paradox. It also enabled a shift from correct logical thinking under wrong preconditions into multiple, agent- and context-based thinking to avoid the logical trap.

The second objection could be that human civilization has not yet accepted the explanation provided here, and the validity of the hypothesis relies on future acceptance of the explanation. The danger is that humans will ignore or oppose the explanation provided here. If so, consequent disclaimers will have to be published in this journal as well. On the other hand, this is the purpose of scientific position papers.

Third, the proposed solutions to the analyzed logical puzzle might seem to be just one single event and not

that the human civilization has improved due to advances in AI, ICT, and cognitive science. However, these and similar paradoxes have stirred human imagination for eons and have not yet been satisfactorily resolved, even by brilliant mathematicians and logicians. In addition, these problems are known globally. Therefore, we must rely on new knowledge when providing the explanation in this paper. Furthermore, in order to show superior computing performance of one mechanism over another, it is necessary to show just one task that a certain mechanism can solve and the other cannot. According to the tentative hypothesis presented here, have we not shown how human mental capabilities have increased in recent decades, since an intelligent individual can understand the solution provided herein but the best knowledge among the smartest individuals could not previously?

This new approach has also been used to solve several other paradoxes, such as the blue-eyes paradox and the Pinocchio paradox. Analyses of these paradoxes are being submitted to other journals.

In summary, the explanation of the hanging paradox and the difficulty for human paradox solvers resembles those of the liar paradox before solving it beyond doubt. It turns out that **both paradoxes are not truly paradoxical**; instead, they describe a logical problem in a way that a human using logical methods cannot resolve the problem. Similar to the untrue assumption that a liar can utter a true statement, the unexpected hanging paradox in the prisoner's line of reasoning exploits **two misconceptions**. The first is that a 100 percent accurate prediction for a single event can be uttered more than once (through a vague definition of "surprise") and the second that a conclusion that is valid at one time is also valid during another time span (moving from Friday to Thursday; that is, not accepting the conditions in statement *C*).

Due to the simplicity of the AI-based explanation in this paper, there is no need to provide additional logical, epistemological, or philosophical mechanisms to explain the failure of the prisoner's line of reasoning. There is nothing wrong with inductive reasoning, as long as preconditions are valid.

The hanging paradox is interesting from various perspectives, such as regarding the question of which methods enable successful analysis and explanation. This paper provides an AI-based explanation for humans, while other explanations, such as an explanation or procedure for AI machines to analyze the unexpected hanging paradox, remain a research challenge.

Acknowledgements

The author wishes to thank several members of the Department of Intelligent Systems, particularly Boštjan Kaluža and Mitja Luštrek, for their valuable remarks. Special thanks are also due to Angelo Montanari, Stephen Muggleton, and Eva Černčic for contributions on this and other logic problems. Also, the anonymous reviewers provided several remarks that helped improve the article.

References

[1] Neisser, U. (1997). Rising scores on intelligence tests. *American Scientist* 85, Sigma Xi, 440–7.

[2] Flynn, J. R. (2009). *What Is Intelligence: Beyond the Flynn Effect*. Cambridge, UK: Cambridge University Press.

[3] Computing laws revisited (2013). *Computer* 46/12.

[4] Moore, G.E. (1965). Cramming more components onto integrated circuits. *Electronics Magazine*, 4.

[5] *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (IJCAI'13) (2013). Beijing, China.

[6] Kurzweil, R. (2005). *The Singularity is Near*. New York: Viking Books.

[7] Dean, T. (2009). A review of the Drake equation. *Cosmos Magazine*.

[8] Wolfram A. (2014). http://mathworld.wolfram.com/UnexpectedHangingParadox.html

[9] Unexpected hanging paradox, Wikipedia (2014). https://en.wikipedia.org/w/index.php?title=Unexpected_hanging_paradox&oldid=611543144, June 2014

[10] Chow, T.Y. (1998). The surprise examination or unexpected hanging paradox. *American Mathematical Monthly* 105:41–51.

[11] Sober, E. (1998). To give a surprise exam, use game theory. *Synthese* 115:355–73.

[12] O'Connor, D.J. (1948). Pragmatic paradoxes. *Mind* 57: 358–9.

[13] Beall, J.C., Glanzberg, M. (2013). In Edward N. Zalta, E.N. (eds.), *The Stanford Encyclopedia of Philosophy*.

[14] Prior, A.N. (1976). *Papers in Logic and Ethics*. Duckworth.

[15] Gams, M. (2001). *Weak Intelligence: Through the Principle and Paradox of Multiple Knowledge*. New York: Nova Science Publishers, Inc.

[16] Sorensen, R. A. (1988). *Blindspots*. Oxford, UK: Clarendon Press.

[17] Young, H.P. (2007). The possible and the impossible in multi-agent learning. *Artificial Intelligence* 171/7.

[18] Turner, R.M. (1993). Context-sensitive Reasoning for Autonomous Agents and Cooperative Distributed Problem Solving, In *Proceedings of the IJCAI Workshop on Using Knowledge in its Context*, Chambery, France.