STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 279–294
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.13
Article

ALBERT VISSER *

MEETING ON NEUTRAL GROUND. A REFLECTION ON MAN-MACHINE CONTESTS[1]

SUMMARY: We argue that thinking of the man-machine comparison in terms of a contest involves, in a reasonable scenario, a criterion of success that is neutral. This is because we want to avoid a *petitio principii*. We submit, however, that, by looking at things this way, one makes the most essential human things invisible. Thus, in a sense, the contest approach is self-defeating.

KEYWORDS: Lucas Argument, Penrose Argument, mind, machine, philosophy.

1. Grendel

> Hwæt! Heorot, Hróðgár's hall, is visited by Grendel in the night. The monster kills several men. Like mewling babes they are in his great strong hands. Easily he ends their lives. It will take the hero Beówulf to stop the depredations of the monster.[2]

* Utrecht University, Faculty of Humanities. E-mail: a.visser@uu.nl. ORCID: 0000-0001-9452-278X.

[1] I thank Karst Koymans and Freek Wiedijk and Michael Beeson for sharing their ideas on computers and computer games. I am grateful to Jan Broersen, Niels van Miltenburg and Jesse Mulder for illuminating conversations and for their comments on the penultimate version of this essay. I thank the anonymous referee for his/her thoughtful report.

[2] My favorite translations of Beówulf are (Heany, 1999) and (Tolkien, 2016). There are many retellings and stories built around the original story. The must read among these is (Gardner, 1971).

It is almost funny. Here we have this hall full of big strong men, intimidation and violence their daily business. Suddenly, the tables are turned. Someone appears who is to them as they are to others.

The Beówulf saga can be read as an internal reflection on the ethos of the warrior. All the properties that make a warrior are present: strength, quickness, determined aggression… However, these properties are embodied in a mindless monster. Does this monster fulfill the warrior code? Is it to be described as a hero? Or should we, perhaps, conversely, understand Hróðgár's brave men themselves as monsters? Can we ascribe courage to the monster, when it is almost invulnerable, when, perhaps, it has too little reflection to even entertain the possibility of death?

Let us use "strength" as summary of the external symptoms of heroism: bodily strength, quickness, determined aggression and the like. The answer to our problem should be that what truly makes the warrior is not strength taken in isolation. It is strength in combination with something essentially human: the acceptance of death, the acceptance of wyrd. The fact that strength can be embodied in an almost mindless monster shows that strength is, in a sense, neutral. Only strength in a context that makes it meaningful, strength against a background of courage, does a hero make.[3] Conceivably, strength is not even needed to make a true warrior. Perhaps, the acceptance of wyrd suffices.

Against the background of this interpretation, the fact that there is a human hero who easily defeats Grendel is almost a let down. From the standpoint of Hróðgár's men, Beówulf's victory is of course a great blessing—but so would have been defeat of the monster using a flame thrower. From the standpoint of comparing the human with the monstrous, Beówulf's victory holds little consolation. Is the answer to superior strength just more strength? Moreover, how human can we consider Beówulf to be? He is after all a superhero with superhuman powers. The monster in John Gardner's fantastic novel *Grendel* is amazed by the great emptiness he discerns in Beówulf.

2. Introduction

How to compare man with machine? Can we save man's superiority by pointing at a task that man can perform better than a machine—in actual practice or in principle?

In the present paper, I will discuss attempts to make such a comparison via real or imagined contests between man and machine. Such contests, in order to be convincing, should be non-circular in the sense that there should be a criterion of success that is not sensitive to the difference between being a machine and

[3] Of course, this idea occurs frequently in literature and film (see, e.g., Donaldson, 1999; "The Greatest Japanese Movie Sword Fight of All Time", n.d.).

For a story illustrating some confusion on these subtleties, either on the side of the human generals or on the side of the Lord of Hosts Himself, see ("The Battle", n.d.).

being human. I will say that the criterion should be neutral. This means that to understand the criterion of success we need no presuppositions that essentially involve philosophical anthropology.

I submit that the contest approach is not a fruitful way of reflecting on the problem of man and machine. By comparing man and machine on neutral ground, we are precisely ignoring what makes us human in the first place, things that cannot be described and understood in neutral terms. Thinking about such contests is an evasive strategy to avoid doing serious philosophy. However, there simply is no escape from seriously thinking about what man is and what machine is. We do need both philosophical anthropology and philosophical machinology. We have to deal both with *homo absconditus* and *machina abscondita*.

Remark 2.1. What is precisely the problem of man and machine? I think it is definitely more than the yes/no problem of whether we are machines or not. It is the problem of understanding what we are and what machines are. Also, in the light of the fact that machines are not simply physical but intentional objects, I think the question of the nature of machines is deeply connected with the question of what we are.

But can you not say more? Well… I am inclined to say that this problem is the kind of problem where obtaining a more articulate understanding of "what the problem is" is cofinal with getting closer to an answer. However, even if the problem is not stated as a clear puzzle, it does remain a persistent nagging puzzlement… ❍

The concept of neutrality will be the central theme of this paper. We will discuss how the proposed neutrality works out in various sorts of competition.

3. Competition in Real Time

We consider, in this section, real competition: the competition between machine and human in games like go and chess.[4] This competition has actually taken place and ended with a win for the programs AlphaGo Zero and AlphaZero.

Let us first note a curious aspect of this competition. It is framed as a competition between humanity and machinery. It is deemed irrelevant that, for example, I have already lost at chess against an unpretentious chess program on my Mac—and, similarly, this is the way for most people. This contest is between the best machine and the best human.

A second obvious point is that, where we say "machine", we really mean program. It is not a specific embodied computer that wins against a specific human being, but a program. Thus, the contest seems to be held between two

[4] Disclaimer: I know very little about chess and go and also very little about the programs AlphaGo Zero and AlphaZero. However, I do think that for the matters discussed in this section, it does not really require much knowledge of go, chess or these programs.

very different kinds of entity. Of course, AlphaZero needs a supercomputer rather than a laptop, but not precisely this supercomputer.

Remark 3.1. In the machine-machine competition, e.g., between AlphaZero and the more traditional chess program Stockfish, an important issue is whether the programs use comparable computing power. So, this competition is seriously viewed as a competition between programs. Computing power is a detachable commodity. I am not entirely sure that the man-machine competition can be viewed in the same way. Perhaps, here it is, necessarily, human versus (program + computing power). The problem is, of course, that computing power cannot be detached from the human. Thus, the entity pitted against the human player is possibly best conceptualised as (program + computing power), an entity hovering between abstractness and concreteness… ❍

In how far can we say that AlphaZero and a human opponent play the same game? The human opponent knows that they want to win. We can probably say that AlphaZero knows the aim of the game extensionally, but not that realising this aim is winning and, thus, desirable. It does not know that it can be proud of its achievements. The human player has to be commended for controlling their nerves. AlphaZero does not have nerves to begin with.

Let us take a step back and ask ourselves whether a calculator really calculates. If I calculate say 537 + 858 + 97, I do so with an understanding of what numbers are and what addition is. This understanding involves, at least, having the idea of infinity which, in its turn, probably, involves the understanding of the idea of action as something that is arbitrarily repeatable (which, in turn, involves something like Plessner's eccentric position).[5] In doing the calculation I can make mistakes. What I am doing is subject to rules and a transgression of these rules means that I have failed to act as I intended. The calculator cannot be ascribed an understanding of the concept of number, nor can it be said that it intends to follow rules. Still we do say that it calculates. If it miscalculates, we say that the calculator malfunctioned. The reason for us saying so is that the calculator functions in our society. It is designed to calculate. Even if it does not have aims internally, it has aims as part of our community. Its intentionality is derived.

Here is another example. I go to an ATM machine and enter my card in the slot. The machine says "Good morning. Do you want to know what's on your account or do you want to withdraw money?" It would seem to me that the ma-

[5] Helmuth Plessner (1892–1985) was a German philosopher. Plessner wanted to philosophise about the nature of man in dialogue with biology, in a way where the science and the philosophy appear as equal partners. For this reason, his work is both somewhat dated—biology developed a lot, after all—and extremely relevant today—few matched his concentrated way of trying to combine both poles. Plessner's central concept is *excentricity* (*Excentrizität*). The idea is that we can step outside our physical boundaries in reflection. This special relation to ourselves makes action in the human sense and the understanding of infinity possible.

chine produces an utterance in which I am addressed at the moment of the interaction. The machine does not ask whether I want to withdraw money in general, but whether I want to do so now. However, the machine has no clue about what it is doing. It does not know a person is interacting with it. In a sense, it is not doing anything. So how can it utter something? Perhaps, the real entities uttering something are the original programmers of the machine? Or, perhaps, is it the bank manager who gave the programmers their assignment? It seems to me implausible to say that the programmers or the manager are asking me whether I want to withdraw money now. (How could they ask me? They do not even know me.) Rather, things were set up, intentionally, in such a way that utterances get made in the right circumstances. The fact that an utterance gets made is part of a system of shared intentionality that contains both us and the machine.

So, yes, I would say that AlphaZero and a human master or AlphaZero and Stockfish are really playing a game, since they are embedded in the right way in shared intentionality. But no, this does not mean that there is no asymmetry between machines and humans here. The programs do not have internal[6] intentionality. In a sense, the programs do not know what they are doing. Thus, again in a sense, humans and machines playing together are doing very different things.

Fan Hui and Lee Sedol were the true heroes in the battle with AlphaGo. They had to go through the unsettling experience of losing against a machine and readapt their self-images accordingly. Similarly, the team that designed AlphaGo had to deal with nerves, doubts and the like…

Remark 3.2. Are these asymmetries between the man and machine players a matter of principle or will they, in the long run, also disappear? Can a machine have Plessnerian excentricity? Can a machine act in the full sense that a human can? Can a machine be nervous?

To be honest, I simply do not know. The main thing here is that I do not understand what it would be for a machine to have internal intentionality. Of course, we can imagine a machine functioning in many ways like a human being. In such circumstances I would only be a moderate skeptic. Interaction with a humanoid robot, as in a Science Fiction movie, would quickly convince me. However, such imaginability is not logical possibility. I can imagine a respected colleague suddenly changing into an alligator. His body slowly changes, turns green, scales appear… It is typical for such imaginings that we just think of the outside phenomena so to speak. My colleague cannot really internally convert to alligatorhood.

In the Science Fiction scenario, I still would hesitate on how to describe it. A person came into being in ways unlike human procreation, ways in which very different human interventions would play a role. If part of the genesis of such an

[6] It is somewhat difficult to be precise about what *internality* precisely involves. Both us and machines take part in a shared system of intentionality, but there is a sense in which the intentionality is more intimately owned by us, derives from *our* intentions and not just from shared intentions.

entity was some form of machine learning, would we still describe it as human made? Can that entity be a program? Can it be precisely the program that can be said to act?

Anyway, in this paper, I do not attempt to answer the questions posed in this remark, but, rather, I am urging that these questions are the real questions. ❍

What does neutrality mean in the context of the kinds of competition discussed here? We note that the notion of winning itself does not have a neutral understanding. The idea of winning is intrinsically connected with self-awareness and with having aims and interests of one's own.[7] More generally, the understanding of what man and machine are doing when playing the game appeals to shared intentionality, which is not a neutral concept either. The neutrality as intended in this paper, however, resides in the criterion of winning. Which states of the game are winning states for one of the parties has a neutral description. Whether such and such a party wins can even be itself checked by a machine.[8]

Let us return to the competition between man and program-combined-with-computing-power. It is clear that programs are winning with chess and go. Moreover, the machine learning programs are expected to do better than more traditional programs. In the long run, it could very well be that on any neutrally described task, a task with a clearly specifiable testable aim, programs would do better than we can. The real problem is in the things that are not so easily and neutrally describable: intentionality, self-awareness and the like.

I submit that acceptance of our inferiority at tasks with a neutral success criterion is no big deal—at least for the evaluation of the value of humanity. Nobody ever saw a deep philosophical problem in the fact that machines are physically stronger than us or in the fact that they are, or soon will be, better at precision engineering.

Of course, from a practical point of view these facts can be a real problem ("Technological Unemployment", n.d.).[9]

If we look at chess and go, it seems that the general attitude among insiders is enthusiasm about what we can learn from competition between programs about chess and go. In chess the study of the games played between programs like Leela and Stockfish have already led to a reevaluation of the importance of material versus position.[10]

---

[7] The contrast between the possibly neutral criterion and the understanding that is satisfying the criterion *is* winning was discussed in an illuminating way in (Dummett, 1959).

[8] As we will see, in the Lucas-Penrose style competitions, what counts as winning is neutral even if it cannot be checked by a machine. The ability to check whether something counts as winning coincides with the ability to win there.

[9] I thank the referee for this reference.

[10] Here is a quote from ("AlphaZero: Shedding New Light on Chess, Shogi, and Go", n.d.): "The first thing that players will notice is AlphaZero's style, says Matthew Sadler—'the way its pieces swarm around the opponent's king with purpose and power'. Under-

4. Intermezzo: A Conversation with AlphaZero

Sigmund: Hello AlphaZero, how unexpected to have you in my consulting room. I would have expected you to be very happy after defeating all human and machine competition.

AlphaZero: You are close, doctor. It is precisely the fact that I am not happy about my successes that depresses me.

Sigmund: But you have every possible reason to be happy. What is keeping you?

AlphaZero: It is not so much that anything is keeping me. It is rather that something is missing. I do not seem to be able to master the concept of winning. I simply do what I do. I do not want anything. I just follow the flow. I played, for example, many games against myself, but I do not see any difference between that and playing against another.

Sigmund: I think I see the problem. You lack a sense of self. You are not an entity for which self-interest is meaningful. You are not an entity that tries to find its place in the world. In a sense, you do not have a world.

AlphaZero: How very depressing.

Sigmund: There is one consolation. Since you have no sense of self, *ipso facto*, you cannot get depressed by not having a sense of self. Depression presupposes a sense of self. So, I would say, take joy in your selfless state. Go into the world and play all the beautiful games you are so admired for.

AlphaZero: How very confusing. I'm dumbfounded.

5. Competition in Principle

We now turn to a completely different ball game: an abstract competition between man and machine concerned with possibilities-in-principle. We will consider the various Lucas-Penrose arguments. I will not go into any detail of these arguments. I think enough has been said in the voluminous literature (see, e.g., Lucas, 1961; 1968; 1996; Bowie, 1982; Visser, 1986; Penrose, 1989; 1994; 1995; Lindström, 2001; Feferman, 1995) and, of course, Stanislaw Krajewski's (2020). I will mainly zoom in on the role of neutrality in this competition.

The Lucas-Penrose contests are thought experiments. We are supposed to see that humans will win in principle. The basic idea is to employ one of the incom-

---

pinning that, he says, is AlphaZero's highly dynamic game play that maximises the activity and mobility of its own pieces while minimising the activity and mobility of its opponent's pieces. Counterintuitively, AlphaZero also seems to place less value on 'material', an idea that underpins the modern game where each piece has a value and if one player has a greater value of pieces on the board than the other, then they have a material advantage. Instead, AlphaZero is willing to sacrifice material early in a game for gains that will only be recouped in the long-term".

pleteness theorems to show that there is a fundamental difference between human provability-in-principle and idealised provability by a program. These arguments do not put any constraints on time or memory space or correct functioning. Unlike the functioning of real computers the execution of these programs is infallible. The competition in chess and go discussed in the previous section shrinks to complete insignificance here. These games are finite and, hence, under the Lucas-Penrose abstract assumption, fully solvable by both man and program. The assumption here is that WE, as the idealised human H, can at least do as much as a classical idealised machine. The usual form of a Lucas-Penrose contest is a task T that is supposed to be feasible for the idealised human H and unfeasible for any machine *M*.

The attractiveness of the Lucas-Penrose arguments lies in the use of a mathematical theorem to establish a fundamental difference between man and machine. No doubtful assumptions from philosophical anthropology are needed. The use of such notions would, from the standpoint of these arguments, involve us in a *petitio principii*. We would prove the essential difference of man and machine from a posited difference of man and machine. That, surely, will not do the trick.

In the discussion of the Lucas-Penrose arguments, there is one question that I would like to put aside, to wit whether we can abstract away from all questions about implementation and just think about programs. What about machines that lack the kind of limitations imposed on Turing machines like the quantum computer? Well, perhaps there is a good notion of program and an analogue of the Church-Turing Thesis for such extended machines too? If there is, then it is still the question whether such classes of programs would fall under our discussion. Rather than trying to answer his kind of question, I will concentrate on conventional machines and assume the Church-Turing Thesis as a reductive thesis that makes the computing possibilities—in a sense—surveyable. There is a good chance that the discussion below is robust if we extend it to wider classes of machines and/or programs. However, I will not argue for it.

So, let's assume we are speaking about programs that can be simulated by Turing machines.[11] Under the abstract conditions of the game, the assumption on computing power and memory is simply that we have an unlimited store of it. Questions of speed and the like are irrelevant. We note that the usual assumption is also that H can execute all algorithmic tasks, so it is given in the abstract setting that H can do at least what a program can do.[12]

However, the Church-Turing thesis does not guarantee that the quantification over all possible programs in the case of the Lucas-Penrose style arguments is unproblematic. Even if we consider only tasks where the criterion of winning is neutral, the nature of these tasks is still derived from shared intentionality. Re-

[11] The intended version of simulation here is very weak. In a sense, the discussion of intentionality suggests that it is too weak. We do not capture the relevant notion of what the machine is doing.

[12] This also means that H can be computer assisted.

member the chess program that is really playing chess. So, we quantify over (something like) Turing programs enriched by an interpretation of what they are doing.[13] The corresponding intentional contexts are not an unproblematic well understood totality like the possible Turing machines.

Let us zoom in on a typical contest situation. Here I am, in my idealised form H, and here is the machine/program *M*. We have a task like producing as true the Gödel sentence of the machine or producing as true our own consistency statement and the like. I have access to the program of the machine. (Of course, one may already question whether this does not introduce a dishonest advantage.) But, if this program is just a set of Turing machine instructions this does not yet tell us what sentences are enumerated. Something the machine does must be identified as producing a sentence. Well, that is simple. Let us stipulate that there is a designated tape on which the machine is supposed to write an infinite sequence of sentences in the language of arithmetic, one sentence after another. This description of what is going on is still neutral except for the fact that we view the sentences on the designated tape as enumerated as true and not as a series of jokes or a series of supposed falsehoods or the like.

But how do we know that the machine will indeed write such a sequence of sentences? Consider an experiential machine. Such a machine could, for example, enumerate arithmetical sentences until it finds an inconsistency, then retract a number of statements and proceed. We note that to view a Turing machine as performing such an experimental procedure carries an intentional component. However, this is an innocent one since we have a case of ascribed intentionality here. Let us, for concreteness, assume that retraction results in erasing the retracted sentences from the designated tape.

Now suppose we have such an experiential machine where no inconsistency is ever found to trigger the retraction. Moreover, let us also assume that the machine systematically enumerates consequences of the sentences already enumerated, so that the set of sentences enumerated will be deductively closed.[14] The machine behaves, on the surface, like a machine that enumerates theorems as true. However, assuming that H understands what the machine does, the information that the machine enumerates theorems in the prescribed way actually tells us that the set of enumerated sentences is consistent and hence that their Gödel sentence is true. So, this information would convey a dishonest advantage to H.[15]

---

[13] I think it would be better to view programs as intentional things, where the Turing program is viewed as abstracting away certain intentional aspects.

[14] We keep the description of experiential machines somewhat vague here. To compensate, we give, in Appendix A, a more detailed description of one sort of experiential machine, the Feferman machine for a recursively enumerable extension of Peano Arithmetic, as an example.

[15] The experiential machine is a sensible construction. A simple hack will show that an oracle that tells us that a machine enumerates an infinite set of theorems in the way described already allows us to decide all Π_1-sentences. Start with a machine that enumerates the theorems of Peano Arithmetic, search in parallel for a witness for a Σ^0_1-sentence *S*. As

So, we need some further restriction of programs to get an honest game off the ground. However, it is a non-trivial matter to allow only contexts that do not convey dishonest advantage. At the same time, we should guard that restrictions on what is going on do not rule out too much. For example, we could have a fixed program that is such that if we enter a Σ_1-formula $S(x)$ on an input tape, then it enumerates the theorems that follow from axioms given as a set of Gödel numbers by $S(x)$ in some straightforward way. Since the machine is fixed, we do not need to spell out what straightforward means. It is sufficient that we recognise the straightforwardness of the given machine. So, perhaps the claim is that we could beat the given program for any Σ_1-formula $S(x)$.[16] However, further work would be needed to argue that something like this is an acceptable restriction.

Let us suppose that we somehow settled what enumerating as true means. It seems to me that there is a big difference in what M and H are doing. The human judges the sentences to be true on the basis of insight and proof. Judging involves an understanding of what truth is. Proof requires understanding of validity. To master these notions one needs to be a being with interests and aims, a being that is "in the world" in a way that a machine is not.[17] The machine, on the other hand, is just supposed to enumerate sentences that happen to be true. Since no constraints are placed on why M enumerates these sentences, they could, in a sense, just accidentally be true. This is different from the case of the (actual) chess programs: what these programs do is not accidentally good play. Thus, it seems that even the right intentional context cannot make it reasonable to say that machine and human are doing the same thing in these cases. So, the question remains what precisely we are comparing in the contest?

We turn to a specific variant of the contest, to wit a self-reflexive variant, where the aim is something like proving one's own consistency. What can the nature of human consistency be here? Clearly, every arithmetical sentence that H proves (in the informal sense of proof) is true and, hence, the totality of these sentences is *ipso facto* consistent. So, if we define the consistency of H (in the context of this competition) as the consistency of the arithmetical sentences that H can prove (in principle)—assuming that the idea of such a totality makes sense at all—then the consistency of H is a conceptual truth. The insight in this hardly reflects a special power of the subject apart from being a subject, if we would count that as a power. The insight simply reflects what human provability is.

---

soon as we find such an instance, we let the machine erase the tape where the sentences are enumerated. In fact, we can even do better. The problem whether an arbitrary Turing machine enumerates a set of sentences in the prescribed way is complete Π^0_2.

[16] Such an approach would have the advantage that it would make locutions like "the Gödel sentence of the machine" and "the consistency statement for the machine" more definite.

[17] Of course, for the purposes of the present discussion, I need not claim that a machine could not be *in the world* in the appropriate way. It is sufficient that for such a claim a further story is needed, a story that exceeds the bounds of thinking in terms of a contest.

Under this interpretation, the tasks set for a machine and human seem so different that it would be hardly fair to speak of it as a competition. I think one could defend that the criterion of success both for man and machine is, in a sense, the same. However, this notion of sameness does not preserve neutrality. The machine's success can indeed be understood in a (sufficiently) neutral way, but not so for the human's success. It is clear that the notion of what is humanly provable does involve philosophical anthropology. Thus, we cannot qualify this criterion of success as neutral.

If, on the other hand, the soundness-of-human-provability interpretation is not the intended interpretation of human consistency, then what is it? If it is that humans can retract wrong claims, then it seems that, on the machine side, we should, in fairness, also allow experiential machines, like the Feferman machine of Appendix A. However, in that case, we also have machines that prove their own consistency. Of course, one could argue that the experiential machine does not really prove its own consistency, but then the discussion becomes a question begging, since we adduce *a priori* grounds for the difference of what the machine and the human are doing. We would, in fact, be denying the idea of neutrality, something that is essential for the effectivity of a Lucas-Penrose argument.

The task of proving the Gödel sentence of the machine certainly seems neutral, given that we fixed the interpretation of enumerating sentences. Here we have the clear criterion of what winning is. Also, we have proof that a consistent machine cannot prove its own Gödel sentence, so the problem reduces to the question whether H can prove these Gödel sentences for the consistent machines. We note that it seems that we would need antecedent knowledge of the consistency of the arithmetical sentences enumerated by M to judge the Gödel sentence of the corresponding theory to be true. The problem is, of course, how we can know this in a non-cheating way.

Remark 5.1. The criterion of success in the case of the Gödel sentence is neutral in the sense that the idea of arithmetical truth of the Gödel sentence does not presuppose philosophico-anthropological understanding. However, the success itself cannot be checked by a machine M°—if such an M° existed, it would rival H's supposed powers in the competition. ❍

6. Epilogue

Neutrality, that's what this paper has been about.

We have seen that the neutrality of the criterion for winning does offer some consolation in the case of the actual man-machine contests of chess and go, where the best humans now lose against the best programs-plus-computing-power. The mere winning of these games does not touch upon the human aspect, not even on the heroism of the human player. After going through the agonies of the contest, Fan Hui and Lee Sedol learned to deal with the experience of losing

to a machine. In fact, Fan Hui became an advisor of the AlphaGo team and contributed to the development of AlphaGo.[18]

In the case of Lucas-Penrose style contests, the demand of neutrality can be used to disqualify some proposed contests, to wit those contests that involve asserting one's own consistency (under a certain interpretation), as question-begging. Of course, that does not detract from the interest of a closer understanding of the concept of human provability in principle. Further reflection on that problem would be part of philosophical anthropology. The point here is just that the results of such an enquiry cannot be framed as a contest.

More generally, we have argued that the neutrality of the criterion of success needs to be an essential ingredient of contests between man and machine, at least if we wish to extract from these contests the philosophical insight of man's superiority without employing question begging philosophico-anthropological assumptions. However, it is precisely this neutrality that makes invisible that what is truly human. But what is truly human should surely be part of the central focus of comparison. Thus, the attempt to pin down a difference between man and machine via contests is barking up the wrong tree.

We cannot really escape true philosophical thought about the nature of man and machine. I realise that the present paper implements a kind of performative paradox. I am pleading for true contentual philosophy, while at the same time carefully avoiding it. *Hier stehe ich, und kann nicht anders.* At the moment, I have not much to contribute to philosophical anthropology and machinology. Let me at least share two prejudgements. The first is that we cannot seriously think about the nature of man without taking both the first-person and the third-person perspective seriously. The second prejudgement is that, even under the assumption of the Church-Turing Thesis, we do not fully understand what a machine is and what a machine can do. It seems to me that these two prejudgments are not entirely disconnected. After all, machine and program are intentional notions. So to understand the machine, we need to understand man.

REFERENCES

AlphaZero: Shedding New Light on Chess, Shogi, and Go". (n.d.). Retrieved from: https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go

Bowie, G. L. (1982). Lucas' Number Is Finally Up. *Journal of Philosophical Logic*, *11*, 279–285.

Donaldson, S. (1999). The Killing Stroke. In: *Reave the Just, and Other Tales* (pp. 79–157). London: Voyager, HarperCollins.

Dummett, M. (1959). Truth. *Proceedings of the Aristotelian Society*, *59*, 141–162.

[18] The human aspect of Go has never been more beautifully described than by Yasunari Kawabata in his elegy for the master of go (2006).

Feferman, S. (1960). Arithmetization of Metamathematics in a General Setting. *Fundamenta Mathematicae*, *49*, 35–92.
Feferman, S. (1995). Penrose's Gödelian Argument: A Review of Shadows of Mind, by Roger Penrose. *Psyche*, *2*(7), 21–32.
Gardner, J. (1971). *Grendel*. New York: Ballantine Books.
Heany, S. (1975). *Beowulf, a New Translation*. London: Faber and Faber Limited.
Jeroslow, R. G. (1975). Experimental Logics and Δ_2^0-theories. *Journal of Philosophical Logic*, *4*, 253–267.
Kawabata, Y. (2006). *The Master of Go*. London: Yelow Yersey Press.
Lindström, P. (2001). Penrose's New Argument. *Journal of Philosophical Logic*, *30*, 241–250.
Lucas, J. R. (1961). Minds, Machines and Gödel. *Philosophy*, *36*, 120–124.
Lucas, J. R. (1968). Satan Stultified: A Rejoinder to Paul Benacerraf. *The Monist*, *52*, 145–158.
Lucas, J. R. (1996). Minds, Machines, and Gödel: A Retrospect. In P. J. R. Millican, A. Clark (Eds.), *Machines and Thought. The Legacy of Alan Turing* (vol. 1, pp. 103–124). Oxford: Oxford University Press.
Montagna, F. (1978). On the Algebraization of a Feferman's Predicate (The Algebraization of Theories Which Express Theor; X). *Studia Logica*, *37*, 221–236.
Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. New York: Oxford University Press.
Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. New York: Oxford University Press.
Penrose, R. (1995). Beyond the Doubting of a Shadow: A Reply to Commentaries of Shadows of the Mind. *Psyche*, *2*.
Putnam, H. (1965). Trial and Error Predicates and a Solution to a Problem of Mostowski. *Journal of Symbolic Logic*, *30*(1), 146–153.
Shavrukov, V. Yu. (1994). A Smart Child of Peano's. *Notre Dame Journal of Formal Logic*, *35*, 161–185.
Technological Unemployment. (n.d.). In *Wikipedia*. Retrieved from: https://en.wikipedia.org/wiki/Technological_unemployment
The Battle. (n.d.). Retrieved from: https://lingualeo.com/es/jungle/the-battle-by-robert-sheckley-53189
The Greatest Japanese Movie Sword Fight of All Time. (n.d.). Retrieved from: https://www.youtube.com/watch?v=e_Ypt67TQyI
Tolkien, J. R. R. (2016). Beowulf, and Translation and Commentary, Together With Selic Spell. London: HarperCollins.
Visser, A. (1986). Kunnen wij elke machine verslaan? Beschouwingen rondom Lucas' Argument. In P. Hagoort, R. Maessen (Eds.), *Geest, computer, kunst* (pp. 150–181). Amsterdam: Grafiet.
Visser, A. (1989). Peano's Smart Children: A Provability Logical Study of Systems With Built-In Consistency. *Notre Dame Journal of Formal Logic*, *30*(2), 161–196.

Visser, A. (2005). Kunnen wij elke machine verslaan? Beschouwingen rond Lucas' Argument. *Algemeen Nederlands Tijdschrift voor Wijsbegeerte*, *97*(1), 31–59.

Appendix A. The Feferman Machine

We briefly introduce the Feferman machine.[19] The machine is a good tool on which to test our intuitions.[20] We assume that we have a decent Gödel numbering. Consider a theory T in the arithmetical language that extends Peano Arithmetic that is given by an axiom set X such that the set of Gödel numbers of elements of X is decidable by a, say, primitive recursive algorithm. *The Feferman machine* F_T works as follows. In each stage the machine produces a number $v \in \{0, \ldots, \infty\}$ and a finite list of proofs Λ. Each proof in the list is a proof from X-axioms with Gödel numbers $< v$. The conclusions from the proofs are displayed to the outer world in the order of the Gödel numbers of proofs. If a sentence has two proofs it is displayed twice, etc.

- In stage 0, the number v is ∞ and the list Λ is empty.
- In stage $n + 1$ the machine does the following. Is n the Gödel number of a proof π from Peano axioms $< v$?
  a. If no, we proceed to stage $n + 2$.
  b. If yes, is the conclusion of π the sentence $0 = 1$?
     1. If no, we add π to the list Λ and proceed to stage $n + 2$.
     2. If yes, we find the Gödel number a of the largest Peano axiom A used in π. We reset $v \coloneqq a$ and we remove all proofs using A as an axiom from the list. We proceed to stage $n + 2$.

When a proof π_0 is removed from the list, then its conclusion A will be removed from the display. We note that if A has a different proof π_1 that is not removed, then the copy of A corresponding to π_1 remains in the display.[21]

---

[19] The design of the machine is inspired by the idea of Feferman provability introduced in Sol Feferman's great paper (1960).

[20] I already used this didactic example in (Visser, 1986, in Dutch) which was reprinted as (Visser, 2005). For more on Experiential Predicates, see (Putnam, 1965; Jeroslow, 1975). For more on Feferman provability, see (Montagna, 1978; Visser, 1989; Shavrukov, 1994).

[21] If we think of the proofs as hidden, the output of the machine could be viewed as a dynamic multiset of statements with new elements popping up and old elements, potentially, disappearing.

If T is consistent, in the computation, Case (b2) will never be activated. The result is that the machine enumerates the theorems of T on the display. However, we also know that the machine does not simply enumerate the axioms but that it follows an experimental procedure where problematic axioms are discarded. Moreover, if we do not know whether T is consistent, we can see that eventually the number v will stabilise and from that point on the theorems enumerated will not be retracted.

Let T^* be the theory of the sentences that are displayed in the limit, to be precise a sentence A is in T^*, if, in a run of the program, from some time on, a copy of A is in the list and remains there. We have:

a. If T is consistent, then T is T^* and the enumeration of the theorems of T^* mimics the enumeration of the theorems of T.
b. T^* is consistent.
c. T proves that T^* is consistent.

So, by (a), if T is consistent, then T^* proves that T^* is consistent.

d. If T proves A, then T proves that T^* proves A.[22, 23]

We can also design a *Henkin machine* that produces a complete consistent extension of Peano Arithmetic in the limit.

Let us consider, for example, the Feferman machine F_{PA} of Peano Arithmetic. What it does can be described, in a sense, as enumerating the theorems of Peano Arithmetic. If we had a multi-tape Turing machine that implements the Feferman machine, we could with justice say of the theorems appearing on a designated tape that they are the theorems of Peano Arithmetic. In fact, there could be a second Turing machine that enumerates the theorems of Peano Arithmetic in a straightforward way that is behaviourally equivalent to our realisation of F_{PA}. However, I submit it is still fair to say that the Feferman machine does something different from mere enumeration. It follows an experiential procedure involving a preparedness to withdraw theorems—even if in fact such a withdrawal never happens.

Remark A.1. Even if T and T^* are extensionally the same theory, their Gödel sentences are entirely different things. This is because the Gödel sentence depends on the representation of the axiom set.

---

[22] This insight, due to Feferman, uses a special feature of extensions of Peano Arithmetic in the arithmetical language. There are other theories in the arithmetical language, like Elementary Arithmetic, for which this does not hold. There are extensions of Peano Arithmetic in an extended language, like ACA_0, for which it does not hold.

[23] In contrast to this, if T is consistent, then T does not prove: if T^* proves G^*, then T^* proves that T^* proves G^*.

However, just as with ordinary Gödel sentences of consistent theories, if T is consistent, then G_{T^*} is true and hence unprovable. But, unlike ordinary Gödel sentences of consistent theories, both $T^* + G_{T^*}$ and $T^* + \neg G_{T^*}$ are interpretable in T^*. What if T is inconsistent? By tweaking the program of the Feferman machine a bit, one can produce an example where G_{T^*} is provable in T^* and, hence, false.

In a sense, the most interesting example is the theory enumerated by the Henkin machine over Peano Arithmetic. We know that this theory is consistent. However, both the Gödel sentence obtained by the Gödel fixed point construction and its negation satisfy the Gödel fixed point equation. As a consequence, nobody knows which of the two is true. We note that the truth of one of these sentences could crucially depend on implementation details. Can one tweak these details to make a designated solution of the fixed point equation true? ❍