

## Assignment VI “Key:” Some Comments on A.L.I.C.E. & Turing

William Kallfelz, grader, PHIL280

Despite some anachronistic passages, the *conceptual* points that Alan Turing (1950) makes are remarkably contemporary. The essential point he makes, in which he more or less spells out his functional criterion, is found in the following passage:

It was suggested tentatively that the question, ‘Can machines think?’ should be replaced by ‘Are there imaginable digital computers which would do well in the imitation game?’ If we wish we make this superficially more general and ask ‘Are there discrete state machines which would do well?’ But in view of the universality property we see that either of these questions is equivalent to this, ‘Let us fix our attention on one particular digital computer *C*. Is it true that by modifying this computer to have adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme, *C* can be made to play satisfactorily the part of A in the imitation game, the part of B being taken by a man?’ (442)

In such a game Turing conjures, an interrogator must determine the sexes of players A and B who communicate with the interrogator via ‘teletype’ (keyboard) interface, and initially are assigned roles played by actual male and female human participants. The interrogator must ascertain the correct sex of the (role) players A and B, with B respectively acting to assist the interrogator, while A thwarting him or her. “We...ask the question, ‘What will happen when a machine takes the part of A in the game?’ Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, ‘Can machines think?’” (434)

Turing admits that there is a latent kind of ‘species/software chauvinism’<sup>1</sup> associated with ‘intelligence’ in setting up the issue in terms of how well the computer *C* ‘s role-playing could dupe an interrogator:

It might be urged that when the playing the ‘imitation game’ the best strategy for the machine may possibly be something other than imitation of the behaviour of a man...I think it is unlikely that there is any great effect of this kind. *In any case there is no intention to investigate here the theory of the game*, [so] it will be assumed that the best strategy is to try to provide answers that would naturally be given by a man. (435, italics added)

Turing, in other words, seems to concede in the italicized passage above entertaining logical possibilities of systems exhibiting a software architecture far different from that characterizable by typically human mental architecture (see note 1 below), but, he goes on to say, the ‘best strategy’ would be to focus on systems constituted by an architecture in the neighborhood of our own. For example, Dreyfus (1999, xlii) discusses the *circularity of relevance* problem, i.e., that of implementing software/hardware capable of generalization along the relatively unique lines of domain-specificity shared by human counterparts. Certainly, as a counterinstance, it would be difficult to engage in a long-term productive dialogue with a system comprised by neural net software that could only recognize patterns attributed as tanks and artillery during cloudy days, but (falsely) attributed as trees and bushes during sunny days!<sup>2</sup>

---

<sup>1</sup> Recall lecture April 26, 2006. In other words, we would tend to recognize ‘intelligent behavior’ in an entity with a ‘mental software architecture’ more or less similar to our own. This would include its possessing: a) A mental ‘dictionary’ or storehouse of a vast ‘background’ lexicon of facts/knowledge often attributed as the folk-notion of common sense. (Something, for instance, consisting of the aim of the CyC project for the last several decades – to build such a ‘background’ repository—a project Dreyfus (1999, xxvii) argues is essentially ill-conceived and misguided.) b) An ‘everyday mind’ consisting of a model of the real world more or less isomorphic to ours. c) A processing module capable of deductive, inductive, and maybe even abductive (or creative association in a particular context, i.e., theory-generation) inference. d) Memory modules (long term and short term) resembling, more or less, the hierarchy of formation into ‘chunks’ of ‘meaningful units’ characteristic of humans.

<sup>2</sup> An actual DoD (Department of Defense) case study Dreyfus cites.

Aside from the tacit software-chauvinism admission, Turing admits only digital-electronic computers into the discussion<sup>3</sup>, because, among other things, of their implementation of a storage, executive, and control units (analogous to the memory, processing, decision/execution aspect attributable to human mental architecture, see note 1 below.) Moreover, because such digital devices “can mimic any discrete state machine...they are *universal* machines.” (441) In other words, Turing equivocates a notion of ‘universal’ here (with the digital implementation of a finite-state automaton) that should *not* be confused with the more general and abstract notion of a UTM. The latter is a paradigm of *interpreter* programs, i.e., programs that run on other descriptions of programs + their associated inputs.

Turing also runs through many objections to the main question posed in his essay, in §6 (442-454):

1. **Theological:** Mentality is a property of souls, and soul-hood is a property endowed to humanity by the creator. Counterargument: “The arbitrary character of this orthodox view,” aside from its obviously speculative and historically incorrect track-record, contradicts itself in placing such restrictions on the theological notion of omnipotence. “In attempting to construct such machines ...rather are we [conversely]...instruments of His will providing mansions for the souls that He creates [?]” (443)
2. **Denial:** Simply, by fiat, declare the possibility (of machines that think) untenable, as the consequences of such “would be too dreadful.” (Counterargument: “We like to believe that Man is in some subtle way superior to the rest of creation.” (444) Like 1., this is a manifestation of wishful thinking.)
3. **Mathematical Objection:** Inherent limitations of any deductive systems, whether expressed theoretically by Goedel or practically in the Halting Problem, apply to digital machines (deemed as universal, in the practical sense, as explained above.) Hence “there will be some questions [in the imitation game] to which it will either give a wrong answer, or fail to give an answer at all however much time is allowed for a reply.” (444). Counterargument: (a) No argument has been given why such limitation-theorems would *not* apply to human intelligence. (445)) b) “[O]ur superiority can only be felt on such an occasion in relation to the one machine over which we have scored our petty triumph. There would be no question of triumphing simultaneously over *all* machines.” (445) **Comment:** The implied fallacy of composition that b) alludes to, however, runs afoul of Turing’s previous claims of *universality*. In other words, assuming in the existence of digital computer *C* as implementing a universal state machine, then, stumping *C* in some self-reference + negation question/answer line of reasoning *would* have implications across *all* machines of that given *type*, since *C* is that *type*’s *representative*.
4. **Consciousness:** Can machines essentially experience qualia? (As alluded to in the citation of Lister’s Oration) Turing counters by conceding that such an objection would pose as “a denial of the validity of our test.” On the other hand, Turing presents a skeptical slippery-slope objection: One could take the extreme position of solipsism and argue that only oneself experiences qualia and conscious agency. “Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks” (446) which among other things renders valid functional criteria of intelligence as posed by the imitation game. Humans can play that game too, after all.
5. **Disabilities:** While admitting that digital machines possess vastly superior computational skills, there are other domains of human mentality (morality, humor, making certain kinds of mistakes, being the subject of one’s own thoughts, etc.) digital machines lack. Counterargument: By ‘scientific induction’ Turing is merely (implicitly) bringing up the Humean objection that it is unwarranted to generalize to all possible cases from one’s experiencing of some limited particulars. (Though ‘by custom’ we do it all the time.) It does not follow, in other words, that no digital machine in the future will ever exhibit such aforementioned characteristics of human mentality. **Comment:** Perhaps Turing’s flimsiest counter-objection concerns the nature of self-reference, i.e., whether or not

---

<sup>3</sup> “[T]he present interest in ‘thinking machines’ has been aroused by a particular kind of machine, ...an ‘electronic computer’ or ‘digital computer.’...we only permit digital computers to take part in our game.” (436)

machines ‘can be the subject of their own thoughts.’ He essentially offers a thin notion of thought-content, which he equivocates with contents of a program. He points out, that given the existence of interpreter-software (implemented as we’ve seen, by UTMs) which can debug or correct other software, we have, in principle, an implementation of a machine “observing the results of its own behaviour [in the sense that] it can modify its own programmes so as to achieve some purpose more effectively.” (449) But interpreter software may differ fundamentally in kind, not just in degree, from the reflective capacity, and other ‘hard problems’ associated with *intentionality*, (about-ness) as discussed in contemporary philosophy of mind and cognitive science, that are so characteristically idiosyncratic of human mentality.

6. **Lady Lovelace’s Objection:** An objection that regardless how sophisticated, *C* in the final analysis can only perform what it is instructed to do, by some outside programmer/agent *A*. Counterargument: The objection is based on the ambiguity inherent in notions of ‘originality’ and ‘creativity.’ In the case of a human being, for instance, “[w]ho can be certain that ‘original work’ that [s/]he has done was not simply the growth of the seed planted in him [or her] by teaching, or by the effect of following well-known general principles.” (450) **Comment:** This objection is perhaps more refined by Dreyfus’ characterization of GOFAI’s (‘good old fashioned AI’s) misguided attempts to create giant databases in the hopes that such huge abstract systems will eventually exhibit the characteristically relative novelty and flexibility of human cognition. (Turing is quite right to point out that such novelty and ‘originality’ never occur in a vacuum.) To this extent, Dreyfus goes on to say, ‘Heideggerian AI’ (really an engineering-robotics research program) exhibits greater promise insofar as connectionist devices are thrust into heuristic environments in which they ‘learn,’ which according to Dreyfus is more akin to the paradigm of human embodied sentience.
7. **Argument from continuity of the nervous system:** The human nervous system is profoundly dissimilar to a digital system. The former, for instance, exhibits some irreducibly analog features its complex firing mechanisms. Counterargument: Though exhibiting continuous-analog features, sufficiently abstracted, an analog differential analyzer can exhibit the same properties as a fully digital machine. (451-452)
8. **Informality of Behavior:** “If each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines.” (452) Counterargument: This is an ‘undistributed middle’ or either/or fallacy. It rests on the equivocation between notions of ‘complete laws of conduct’ versus ‘complete laws of behavior.’ For example, in contemporary terms, one distinguishes between semantics and pragmatics. No one would claim that even if in principle there could exist a finite and complete semantic lexicon (as in the case of linguistic competence, [Chomski]), this somehow would determine pragmatics (linguistic performance), which is obviously open-ended.
9. **ESP objection:** See Theological objection

Having laid out these nine objections, Turing positively claims that machines with software could be designed modeling the cognitive architecture of the minds of children (456), insofar as the possibility is feasible that a number of connections (or degrees of freedom of interconnected information-flow) could form a ‘supercritical’ upper bound, creating the possibility for learning and creative-idea generation. By and large his remarks are speculative, though rather similar in spirit to Dreyfus’ description of the problem of reinforcement-learning heuristics (1999, xli-xlvi) Dreyfus comments, though based on the advantage of historical hindsight, are more guarded and skeptical.

Overall, Turing points out that much of the nine objections above stem from the objection of consciousness (4) or some variant thereof. Regardless how one may respond to Turing’s arguments, if nothing else he surveys the issues with breadth and clarity. Regarding, however, the ‘harder’ problems associated with intentionality (my comment to his ‘disabilities’ objection 5) he glosses over, I myself tried repeating a short conversation with ALICE several times, where each

and every time the program simply ground to a halt. This conversation had to do with getting Alice to reflect upon the contents of the conversations, and specifically on 'her' own responses thereon:

Here is a sample. I begin by asking: 'How are you?' Alice responds 'Everything is going extremely well for me.' Whereupon I ask:

**Human: Why is everything going extremely well for you?**

**ALICE: Why do I exist?**

**Human: Are you exaggerating?**

**ALICE: I don't know whether or not I am exaggerating. I am well respected as an AI.**

**Human: You didn't answer my first question**

**ALICE: Please help me improve my communication skills. What would have been an appropriate response?**

**Human: You said "extremely well," which is unusual to say**

**ALICE: That's not what I meant.**

**You say:**

what did you mean?

Each time I entered this question (based on the above line of interrogation, which I was able to duplicate, the program terminated.)

**References:**

ALICEBOT URL = <<http://www.alicebot.org> >

Dreyfus, Hubert (1999) *What Computers (Still) Can't Do*, MIT Press

Turing, Alan (1950) "Computing Machinery and Intelligence," *Mind* vol 59, n 236, 433-460.