(H3) “Strong” AI/Computationalism Defined

Selmer Bringsjord
Philosophy of AI

September 24, 2001

1 First Part of Computationalism

This handout says that there are three parts to a theory of mind for “aggressive” or “Strong” AI and Cog Sci (= computationalism).

The first part of the theory is a version of functionalism; we’ll call it **AI Functionalism**. Here’s an informal version of this thesis:

First, a picture ...

Next, the thesis in prose:

(AI-F) For every two “brains” x and y, possibly constituted by radically different physical stuff, if the overall flow of information in x and y, represented as a pair of flow charts (or a pair of Turing machines, or a pair of Turing machine diagrams, ...), is the same, then if “associated” with x there is an agent s in mental state S, there is an agent s' “associated” with y which is also in S.

By the phrase ‘mental states’ I mean nothing magical. (AI-F) refers by this locution simply to those states innocently denoted by such gerundives as *Selmer’s fearing ghosts*, *Bush’s worrying about Cuomo*, *Smith’s being sad*, and so on.

2 Second Part of Computationalism

The second part of the theory of mind underlying the “ambitious” version of AI with which we’re concerned is that

(PER-AUT) Persons are automata.

Automata, not machines. The term ‘machine’ connotes concreteness; it’s a term which evokes in many a mental picture of cogs, levers, circuits, and so on.

Automata need not be corporeal: agent dualism (roughly, persons are individual incorporeal entities) is perfectly consistent with (PER-AUT) — though in practice, most AIniks reject agent dualism.

As we proceed in the course, we will refine (PER-AUT), but it should do to get us started.[1]

Given (H2), you know that (PER-AUT) may be made more specific in many ways: 'automaton' in this thesis may for the moment be replaced by anything from 'digital computer,' to 'finite automaton,' to 'infinite abacus,' to 'cellular automaton,' to 'universal Turing machine'.

Since we're on the subject of what persons are, I would be remiss if I didn't mention two general views on the matter, namely, agent dualism and agent materialism.

Agent dualism is the view that persons, or agents, are not physical things. On this view persons are neither identical with their bodies nor a proper part of their bodies. Persons are non-physical entities, like numbers or times or relations or propositions or sets, etc.

Agent materialism is the view that persons, or agents, are physical things, most likely human brains or some proper part of human brains (the neo-cortex, say), along with certain supporting structures.

It's important to note that (PER-AUT) is consistent with agent dualism. This is so because automata are, at bottom, just sets, and sets aren't (usually anyway) considered to be physical things.

On the other hand most AIniks do affirm agent materialism.

Here are two arguments which we will briefly discuss, one for (something at least on the way toward) agent dualism, one against it. I will claim that the "pro" argument is formidable, but that the "con" argument is unsound.

2.1 An Argument For Agent Dualism

1. $\forall x, y$ (x has a y $\rightarrow x \neq y$)
2. For every person S and brain-of-S B: S has a B.
3. For every person S and brain-of-S B: $S \neq B$. (from 1, 2)

Notice that this argument can be instantiated to work with respect to any individual person, as for example in

1. $\forall x, y$ (x has a y $\rightarrow x \neq y$)
2. Selmer has a brain (call it B_S).

[1] One such refinement, as we'll see later on, is the following proposition (which I'll explain later).

$$\forall x\,(Px \wedge x \text{ is conscious from } t_i \text{ to } t_{i+k} \Rightarrow \exists y\,(My \,\wedge\, x = y \wedge C_j \vdash_y C_{j+1} \vdash_y \cdots \vdash_y C_{j+p})),$$

where this computation is identical to the consciousness x enjoys through $[t_i, t_{i+k}]$.

3. Selmer $\neq B_S$. (from 1, 2)

2.2 An Argument Against Agent Dualism

(1) If agent dualism is true, then there is a distinct entity in which reasoning, emotion, and consciousness take place (the agent, or person), and that entity is dependent on the brain for nothing more than sensory experiences as input and volitional executions as output.

(2) If there is a distinct entity in which reasoning, emotion, and consciousness take place, and that entity is dependent on the brain for nothing more than sensory experiences as input and volitional executions as output, then one would expect reason, emotion, and consciousness to be relatively invulnerable to direct control or pathology by manipulation or damage to the brain.

(3) It's not the case that reason, emotion, and consciousness are relatively invulnerable to direct control or pathology by manipulation or damage to the brain (consider alcohol, narcotics, etc.).

(4) Agent dualism is false. (from 1, 2, 3)

I don't think any agent dualist worth his or her salt would agree to the agent-brain relation that is assumed in premises (1) and (2). The agent dualist doesn't have to be a dummy: she can readily admit that the agent can be harmed (etc.) by doing things to the brain; she will certainly be aware of the empirical facts to which this argument appeals, namely that if you cause brain effects through the giving of drugs (say), the agent involved is likely to have all kinds of strange emotions, beliefs, and so on.

I could be challenged to say more about the connection between a person's brain and a person. But the question is, do I *need* to say any more than that this connection an intimate one — one which makes it the case that stimulating (etc.) the brain causes the person to undergo changes? I don't think so. Because consider how the new version of premise (2) would read once my rebuttal is incorporated:

(2$'$) If there is a distinct entity in which reasoning, emotion, and consciousness take place, and that entity is intimately connected to the brain, then one would expect reason, emotion, and consciousness to be relatively invulnerable to direct control or pathology by manipulation or damage to the brain.

While premise (1) would perhaps be true in the new version of the argument, and while this argument would still be valid, (2$'$) would seem to be false.

3 Third Part of Computationalism

The third part of computationalism reflects the engineering aspect of "Strong" AI; it's concerned, generally, with what AI researchers will be able to accomplish.

Two ways of encapsulating what AIniks will be able to accomplish:

(PBP) AIniks will succeed in building persons.

The second is by way of the proposition

(ROB) AIniks will eventually build a robot able to pass stronger and stronger versions of the Turing Test.

Note that (PBP) and (ROB) are independent of each other (neither implies the other). Building persons isn't the same thing as building robots who look just like persons.

One of my objectives will be to argue against (PBP). (I will argue *for* (ROB).) The high-level structure of my attack is pretty straightforward; it's simply this argument:

(5) ¬(PER-AUT)

(6) (PBP) → (PER-AUT)

(7) ∴ ¬(PBP)

As you now know (or at least should know given your assimilation of (H1) "Logic Tools"], this little argument is formally valid, since it's an instance of modus tollens.

Given what we've covered earlier in this handout, there are at least two other sorts of attack on what might be called the Person Building Project:

(8) ¬(AI-F)

(9) (PBP) → (AI-F)

(10) ∴ ¬(PBP)

(11) ¬Agent Materialism

(12) (PBP) → Agent Materialism

(13) ∴ ¬(PBP)

4 An Alternative Formulation of Computationalism

An alternative formulation of Computationalism, one I've used recently in a couple of papers, is as follows.
Computationalism consists of the following four propositions.

CTT A function f is effectively computable if and only if f is Turing-computable.

P=aTM Persons are Turing machines.

TT The Turing Test is valid.

P-BUILD Computationalists will succeed in building persons.

TT-BUILD Computationalists will succeed in building Turing Test-passing artifacts. (This proposition is presumably entailed by its predecessor.)