Belief in the Singularity is Logically Brittle[*]

Selmer Bringsjord
Department of Computer Science
Department of Cognitive Science
Lally School of Management & Technology
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
Contact: Selmer.Bringsjord@gmail.com

March 17 2012

[*] I owe a unpayably large debt to both the guest editor and an anonymous referee for wonderfully insightful questions and comments regarding earlier drafts.

The dominant purportedly rational basis ($\mathcal{A}$) for believing "The Singularity, barring defeaters, will eventually come to pass" (**S**) is seductive when left informal, but exceedingly brittle when even a smidgeon of formal logic is brought to bear.[1] Some natural-language statement S is **logically brittle** if and only if, once S is respectably formalized, either it's provably false on that formalization, or the dominant basis for believing the statement is provably unsound (again, on that formalization). I take it that despite the entertaining narratological gymnastics of (e.g.) H.G. Wells and other fiction writers, the statement (T) "In the future, we will be able to travel back in time and prevent the Holocaust" is brittle. For it may well be the case that, once formalized, T is inconsistent with an accurate axiomatization of physics.[2] Please note that any case for logical brittleness must be hypothetical in nature, for the case succeeds when one shows that *if* the respectable formalization in question is affirmed, then the S in question is undermined. I now demonstrate that $\mathcal{A}$ is logically brittle.

The "rational basis" to which I refer is rooted in the reasoning of Good (1965), ably amplified by Chalmers (2010), and, reproduced here (essentially verbatim) for ease of reference, the following argument.

	$\mathcal{A}$:	
	Premise 1	There will be AI (created by HI and such that AI = HI).
	Premise 2	If there is AI, there will be AI$^+$ (created by AI).
	Premise 3	If there is AI$^+$, there will be AI^{++} (created by AI$^+$).
∴	**S**	There will be AI^{++} (= $\mathcal{S}$ will occur).

To understand the argument, note the following, which follows Chalmers directly. 'AI' is artificial intelligence at the level of, and created by, human persons, 'AI$^+$' artificial intelligence above the level of human persons, and 'AI^{++}' super-intelligence constitutive of $\mathcal{S}$. (I reserve 'FAI' to refer to the *field* of AI.) Moreover, each of these three constants designates a class of **machines**, where each member of a class has a maximum **level** of **intelligence**, and where the key process is the **creation** of one class of machine by another. I've added for convenience 'HI' for human intelligence; the central idea is then: HI will create AI, the latter at the same level of intelligence as the former; AI will create AI$^+$; AI$^+$ will create AI^{++}; with the ascension proceeding *ad indefinitum*.

It all sounds so ... *inevitable*; hence the seduction. But what do these bolded concepts *mean*? Mathematically speaking, what is a machine, what is intelligence (and a level thereof), and what is the process of creation that stands at the heart of the informal yarn that those who take $\mathcal{S}$ seriously spin? Chalmers (2010), despite a welcome gesture in the direction of rigor (pp. 24–26), gives no answers. Fortunately, thanks to formal logic, and its having given birth to rigorous computer science (Halpern et al. 2001), we have more than the standard metaphors available, and more than science fiction as a foundation for judging whether The Singularity is silly or serious. The machines in the dialectic of which the present short note is a part are obviously *information-processing* machines; the intelligence of these machines can hence be respectably formalized as their in/ability to compute certain number-theoretic functions; and the level of the intelligence of a given class of machines can be respectably formalized as the class of such functions these machines can compute. To render

[1] Defeaters, following Chalmers (2010), are such things as natural cataclysms. Please note that **S** is by my design *maximally* temporally latitudinarian. Since, as I show herein, if the math does happen to break against those who predict The Singularity (qua event; $\mathcal{S}$) will occur within some interval (a century, e.g.), the math says $\mathcal{S}$ will *never* occur, period: The amount of time is irrelevant. Of course, herein I aim to show only that the math *could* break against those who predict $\mathcal{S}$ will obtain.

[2] The reduction of physics to formal logic is now well underway, and progressing swimmingly. E.g., see (Andréka, X, Németi & Székely 2008).

this framework concrete and perspicuous for present purposes, we need only consider three machine classes that appear early on in the hierarchy:[3]

- $\mathcal{M}_1$: push-down automata
- $\mathcal{M}_2$: standard Turing machines
- $\mathcal{M}_3$: infinite-time Turing machines

Now, what about the process of creation? That's easy. For a machine M to create a machine M' is for the former to start processing its inputs, carry out some work, and leave as output M'. More formally, if we allow the subscript to pin down the class $\mathcal{M}_i$ in question, and a superscript to simply indicate some particular machine in the relevant class, we as humans can easily enough build some M_2^k that begins its processing with an empty tape, and leaves on that tape a new Turing machine M_2^m at the end of its work. We of course prohibit the use of oracular information; this is regimented by insisting that at the start of processing the answer cannot be pre-loaded on the tape. In general, we can write $M : u \longrightarrow v$ to indicate that M starts its processing with string u on its tape, and concludes with string v there — which allows us to be clear about one machine producing another: We can simply write $\langle M \rangle$ to denote the "stringification" of the machine M. To ease exposition, when we say that a class $\mathcal{M}_i$ of machines creates a class $\mathcal{M}_j$ where $i < j$ and hence that the latter class is more intelligent, we mean that there's a machine in the former class that creates a machine in the latter able to compute a function no machine in the former can compute.

Chalmers informally lists the techniques currently in use by HI in FAI (again, the field of AI) broadly understood: "brain emulation," "artificial evolution," "direct programming," and "machine learning." Under the formal framework adumbrated above, and under the brute empirical fact that official, archival-quality work by HI toward AI is, foundationally and mathematically speaking, firmly activity at and below $\mathcal{M}_2$ (e.g., for confirmation, see the encyclopedic Russell & Norvig 2009), each member of Chalmers' quartet is firmly at or beneath the Turing Limit; that is, firmly at or beneath $\mathcal{M}_2$. Moreover, all the types of intelligent machines (or agents) in FAI are at and below $\mathcal{M}_2$ (for confirmation see: Russell & Norvig 2009). So, to make explicit both the nature of AI, and the techniques available to HI for creating AI, note that part of my formalization is the following pair.[4]

Proposition 1: AI is at the level $\mathcal{M}_2$ or below.

Proposition 2: The processes available to HI for creating AI are all at the level of $\mathcal{M}_2$ or below.

What supports this pair of propositions? Both could in fact be laboriously proved relative to the current specification of the quartet of techniques Chalmers relies upon (recall above), and relative to the specification of machine (or agent) types that, a la (Russell & Norvig 2009) and other definitive reference works (e.g., Luger & Stubblefield 1993), define FAI. Such proofs are of course well beyond the scope of the present note, but how would they work? The most economical path would be to

[3] $\mathcal{M}_1$ and $\mathcal{M}_2$ are covered in any good introduction to computability theory; I recommend (Lewis & Papadimitriou 1981), which also provides elegant coverage of complexity, a topic relevant to the ending of the present note. $\mathcal{M}_3$ is somewhat more advanced, but all that is needed (assuming some background in set theory) can be found in (Hamkins & Lewis 2000).

[4] Here I use 'AI' in a manner that directly follows Chalmers' use of the term to denote a class of information-processing machines (or agents) that are created by HI, and are such as to be created by such techniques as we see in play before us in this day and age; e.g., by "artificial evolution."

first express Chalmers' quartet as a series of algorithms (easily enough done); to then note that each and every agent type specified in FAI and its definitive reference works is expressed as a series of algorithms (indeed, agents are often specified via standard pseudo-code); and to then invoke Church's Thesis in a manner parallel to its standard deployment in proofs in theoretical computer science (e.g., see Lewis & Papadimitriou 1981) — which would allow immediate identification of all the algorithms in question with standard Turing machines, that is, with $\mathcal{M}_2$.

We can now prove via either or both of two routes that **S** is logically brittle. In the first route we have no need of Propositions 1 and 2; in the second we exploit this pair. I take now the first route, which shows that Premises 2 and 3 are false on a respectable formalization (viz., the present one), and hence that the basis $\mathcal{A}$ for **S** is logically brittle:

> **Theorem 1**: Necessarily, $\mathcal{M}_2$ can't create $\mathcal{M}_3$.
>
> **Proof**: Suppose for *reductio* that our target is false; i.e., that $\mathcal{M}_2$ *can* create $\mathcal{M}_3$. We know that no machine in $\mathcal{M}_2$ can solve the famous halting problem (HP). We also know that HP *can* be solved by machines in $\mathcal{M}_3$; let M_3^k be such a machine. Then it follows immediately that there's a machine M_2^n able to solve HP, which is absurd; so by *reductio* we're done. **QED**

Theorem 1 can be generalized in the context of relative computability to: **Theorem 1***: $\forall i, j : \mathcal{M}_i$ can't create $\mathcal{M}_j$, where $i < j$. This more general result obviously falsifies both Premise 2 and Premise 3, since both of these premises claim — under the present formalization — that lower-level-to-higher-level creation will happen, and is hence, contra Theorem 1*, mathematically possible.[5] That is, explicitly, if either Premise 2 or Premise 3 is true, then — under the present formalization — $\exists i, j : i < j \wedge \mathcal{M}_i$ can create $\mathcal{M}_j$.[6] But this existentially quantified formula is provably inconsistent with Theorem 1* using elementary first-order reasoning. Note that neither Proposition 1 nor Proposition 2 are needed.

Formally speaking, the second route to establishing the logical brittleness of $\mathcal{A}$ is overkill, but the route is worth unpacking, because it both takes direct account of both the nature of FAI and Chalmers' understanding of that nature (via Propositions 1 and 2), and because it reflects an elevated view of human intelligence (HI) that (i) isn't without adherents, and (ii) is at any rate formally respectable. For the second route, we begin with:

> **Theorem 2**: If HI = $\mathcal{M}_p$, where $p > 2$, then, given Proposition 1, there will never be AI such that AI = HI; i.e., Premise 1 in $\mathcal{A}$ is false.
>
> **Proof**: Suppose that the antecedent holds, recall that AI is indeed $\mathcal{M}_2$ or below by Proposition 1, and assume for *reductio* that there will be AI such that AI = HI. Given these suppositions, standard Turing machines are able to solve HP (since by identity with HI and the antecedent they can as $\mathcal{M}_3$-or-more-powerful machines solve HP). But that is absurd, and once again by indirect proof we are finished. **QED**

Of course, in order to show that on the formalization in question the second route leads to the unsoundness of $\mathcal{A}$ by way of falsifying Premise 1, one must include in this formalization the antecedent of Theorem 2 (which then by *modus ponens* on Theorem 2 yields the negation of Premise 1 immediately), but that antecedent, while certainly controversial from the standpoint of the ongoing search for "ground truth," is without question formally respectable, as for example

[5] Review of relative computability is outside the scope of the present note. For interested readers, I recommend starting with the gentle treatment of the Arithmetic Hierarchy in (Davis, Sigal & Weyuker 1994).

[6] Momentarily, I end by considering an alternative formalization, viz. one in which AI, AI$^+$, AI^{++}, etc. are all within $\mathcal{M}_2$.

even formidable thinkers like Gödel, writing even before Good, can be shown to have demonstrated (e.g., see the recursive proof in Bringsjord et al. 2006).

But are there *other* respectable formal frameworks in which $\mathcal{A}$ turns out to be sound? Since the argument $\mathcal{A}$ is, whatever else its defects, formally valid, this query distills to: "… formal frameworks in which the premises of $\mathcal{A}$ turn out to be true?"[7] Some will no doubt say Yes, and for the sake of brevity I shall happily concede the affirmative; but I have my doubts. One possible alternative formalization for generating an affirmative response is simply one based entirely in computational complexity applied to standard Turing machines, and below. This would mean that all the machines referenced in the succession at the heart of $\mathcal{A}$ vary only in the *speed* with which they compute Turing-computable functions. Perhaps, then, the idea would be that an ultraintelligent machine is one unfazed by NP-completeness, but incapable of surmounting the Turing Limit. This direction seems bizarre, since it would mean that ultraintelligent machines, despite being ultraintelligent, can't process information in ways that we humans now, courtesy of our mathematical ability, understand, and specify mathematically (witness infinite-time Turing machines, and — see e.g. (Bringsjord & van Heuveln 2003) — proofs that explicitly use infinitary reasoning). Or to put the point another way: Good (1965) tells us: "Thus the first ultraintelligent machine is the last invention that man need ever make." This would be false, clearly, on the complexity-based formal framework intended to validate $\mathcal{A}$. The reason is that now, at this very moment, as I write this final sentence, many of the best and brightest minds falling into the class HI are trying to *invent* ways of implementing the information-processing machines in class $\mathcal{M}_3$ and above.

References

Andréka, H., X, J. M., Németi, I. & Székely, G. (2008), 'Axiomatizing Relativistic Dynamics Without Conservation Postulates', *Studia Logica* **89**(2), 163–186.

Bringsjord, S., Kellett, O., Shilliday, A., Taylor, J., van Heuveln, B., Yang, Y., Baumes, J. & Ross, K. (2006), 'A New Gödelian Argument for Hypercomputing Minds Based on the Busy Beaver Problem', *Applied Mathematics and Computation* **176**, 516–530.

Bringsjord, S. & van Heuveln, B. (2003), 'The Mental Eye Defense of an Infinitized Version of Yablo's Paradox', *Analysis* **63**(1), 61–70.

Chalmers, D. (2010), 'The Singularity: A Philosophical Analysis', *Journal of Consciousness Studies* **17**, 7–65.

Davis, M., Sigal, R. & Weyuker, E. (1994), *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science*, Academic Press, New York, NY.

Good, I. J. (1965), Speculations Concerning the First Ultraintelligent Machines, *in* F. Alt & M. Rubinoff, eds, 'Advances in Computing', Vol. 6, Academic Press, New York, NY, pp. 31–38.

Halpern, J., Harper, R., Immerman, N., Kolaitis, P., Vardi, M. & Vianu, V. (2001), 'On the Unusual Effectiveness of Logic in Computer Science', *The Bulletin of Symbolic Logic* **7**(2), 213–236.

Hamkins, J. D. & Lewis, A. (2000), 'Infinite Time Turing Machines', *Journal of Symbolic Logic* **65**(2), 567–604.

[7] Perspicacious cognoscenti might appeal to a framework in which one gives up the classification of AI machines as falling at or below the Turing Limit, for in this case my proof of Theorem 2 is blocked. However, (i) a variant of the theorem could rely on a formalization of the techniques for creation, which as reflected in Proposition 2 are themselves Turing-level; and at any rate (ii) there remains the problem that (P2) and (P3) are falsified by Theorem 1.

Lewis, H. & Papadimitriou, C. (1981), *Elements of the Theory of Computation*, Prentice Hall, Englewood Cliffs, NJ.

Luger, G. & Stubblefield, W. (1993), *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*, Benjamin Cummings, Redwood, CA.

Russell, S. & Norvig, P. (2009), *Artificial Intelligence: A Modern Approach*, Prentice Hall, Upper Saddle River, NJ. Third edition.