Are We Paraconsistent? On the Luca-Penrose Argument and the Computational Theory of Mind

Jason L. Megill
University of Virginia

Abstract: I argue that if we are Turing machines, as the Computational Theory of Mind (CTM) holds, then we are paraconsistent, i.e. we do not implement classical logic as canonical versions of the CTM generally hold (or assume). I then show that this claim presents a serious challenge to the Lucas-Penrose argument (Lucas 1961, Penrose 1989, 94), as it collapses Lucas-Penrose into a disjunction (in a manner reminiscent of Benacerraf's (1967) famous objection to Lucas-Penrose). Specifically, whereas Lucas-Penrose concludes that we are not Turing machines, I show that the most one can conclude from the argument is that *either* we are not Turing machines *or* we are Turing machines implementing a non-classical logic.

In 'Minds, Machines and Gödel,' J.R. Lucas (1961) put forth an argument against any mechanistic theory of mind that attempts to equate the human brain with a Turing machine (TM). Roughly, Lucas reasoned: (1) no consistent formal system (or TM implementing a formal system) can decide the Gödel sentence ('I am not provable'), (2) the human mind can decide the Gödel sentence (i.e. we can look and see the truth of the sentence), therefore (3) the human mind cannot be a TM. Lucas' argument remains relevant—and continues to generate debate—as: (1) it can be taken as an attack on the enormously influential Computational Theory of Mind (CTM), and (2) Lucas' argument has been revived, defended and expanded in two recent books by R. Penrose (1989, 94).

Here, I argue that if we are in fact TMs, we implement a paraconsistent logic, and *not* a classical logic (FOL), as canonical versions of the CTM generally hold (or implicitly assume). I show that simply raising this possibility is enough to defeat Lucas' attempt to respond to Putnam's (1960, 95) devastating criticism of

Lucas-Penrose (L-P). Even worse for L-P, the possibility that we are paraconsistent entails that the *most* that we can conclude from L-P is that *either* (1) we are not TMs *or* (2) we are able to see the truth of the G. sentence because we are paraconsistent TMs (as a paraconsistent TM would be able to decide the G. sentence). That is, raising the possibility that we are paraconsistent is sufficient to collapse Lucas-Penrose into a disjunction (so this criticism is at least as severe as Benacerraf's (1967) famous objection to L-P, which also collapses L-P into a disjunction).

First, I offer a very brief introduction to paraconsistent logic, before arguing that we are in fact paraconsistent. Then, I show how the possibility that we are paraconsistent serious undermines L-P.

Paraconsistent Logic

Paraconsistent logic – unlike classical logic – denies that anything follows from a contradiction. That is, in paraconsistent logic, the inference from p and $\sim p$ to any q is blocked. Classical systems, if inconsistent, collapse into a useless heap; if any q can be deduced in a system the system is essentially useless. Paraconsistent systems, however, can allow for a contradiction and still remain useful, as the contradiction cannot be used to infer anything. In short, a paraconsistent approach prevents a contradiction from infecting an entire system. (For an introduction to paraconsistent logic see Priest, Routley and Norman (1989), Priest and Tanaka (1996), or Priest forthcoming.)

Relevance logic is the most well-known, developed and discussed type of paraconsistent logic. In addition to rejecting the inference from p and $\sim p$ to any q, relevance logic also rejects the two other 'paradoxes of strict implication' ($(p \rightarrow (q \rightarrow q)$ and $p \rightarrow (q \rightarrow p)$) and the 'paradoxes of material implication' ($(p \rightarrow (q \rightarrow p))$, $\sim p \rightarrow (q \rightarrow q)$ and $(p \rightarrow q) \vee (q \rightarrow r)$). The motivation for relevance logic is a belief that classical logic allows for 'fallacies of relevance.' That is, classical logic unjustly neglects the actual relationship between antecedent and consequent (and premise and conclusion), a circumstance that allows for 'dubious' inferences where the antecedent seems irrelevant to the consequent, such as

'My feet are cold, therefore, if there is someone at the door then my feet are cold.' (For an introduction to relevance logic, see Mares (1998); see Anderson and Belnap (1975) for the canonical formulation of relevance logic.)

Paraconsistent logic has numerous interesting philosophical implications. One such implication is the following: paraconsistent logic would be one way of overcoming the limitations of the formal method when it comes to attempted formalizations of arithmetic insofar as it might be possible to prove the Gödel sentence in a paraconsistent system (see Priest 1996, for example). In a classical system, if the Gödel sentence is provable, then the system is inconsistent, and if the system is consistent, then the Gödel sentence is not provable (though it is true, hence the system is incomplete). But for paraconsistent systems, the threat of inconsistency is no longer a threat, a circumstance that implies that 'the Gödel sentence may well be provable' (Priest 1996) in a paraconsistent formalization.

Are We Paraconsistent?

I wish to raise the following possibility: if we are in fact TMs (as the CTM holds), we do not implement classical logic, as canonical versions of CTM argue or assume; rather, we are paraconsistent. (For the principle formulations and defenses of the CTM, see Putnam (1960, 67), Fodor (1975, 81, 87, 90, 93) and Pylyshyn (1980, 94), for example.) As I now argue, there are several good reasons for suspecting that we utilize paraconsistent reasoning, which implies that if we are in fact TMs, then we implement a paraconsistent logic.

The first issue I discuss concerns the frame problem. Recall Dennett's (1984) diagnosis of the multi-faceted frame problem. A series of robots (R1, R1D1, R2D1) are faced with the following scenario: safely retrieve your batteries from a room in which a ticking bomb is also placed. One of the robots (R1D1) is blown up because it takes too long calculating irrelevant details and implications of its actions, such as 'pulling the wagon out of the room (will) not change the color of the room's walls' (Dennett 1984). That is, one aspect of the frame problem concerns the

question as to how we instinctively seem to focus on the more relevant implications of our acts, and how we can program an 'intelligent' robot to do the same.

Further, recall that relevance logic—the most prominent form of paraconsistent logic—is motivated by a desire to eliminate 'fallacies of relevance.' The implication is: if we do in fact implement some type of relevance logic, then this might explain why *we* don't suffer from this aspect of the frame problem, i.e. this might account for how we naturally ignore irrelevant consequences of our actions when attempting to determine the consequences of our actions. This result should not be underestimated: the frame problem is a tremendous difficulty facing not only Symbolic Artificial Intelligence but cognitive science as well. As Fodor (2000, p. 42) states, 'the frame problem is a lot of what makes cognition so hard to understand...cognitive science minus the frame problem is Hamlet without anybody much except Polonius.' If the possibility that we are paraconsistent can help unravel the frame problem, then the possibility that we are paraconsistent should be taken seriously.

It seems to be a brute fact of our cognitive lives that our reasoning does not generally or systematically contain inferences such as: 'My feet are cold, therefore, if there is someone at the door then my feet are cold.' If our reasoning had no concern for relevance, it is hard to see how we could function properly. Of course, there might be exceptions: it could be argued that sometimes we do make dubious inferences such as 'I won the game because I wore my lucky hat,' but even this example can be interpreted as being a case of misplaced relevance, as opposed to a lack of relevance.

To continue, Priest (see (1996), for instance) points out that there are numerous examples of inconsistent scientific theories. An example is Bohr's theory of the atom: Bohr's theory states that 'an electron orbits the nucleus of the atom without radiating energy,' while Maxwell's equations, which play an important role in Bohr's theory, state that 'an electron which is accelerating in orbit must radiate energy' (Priest 1996). As 'not everything concerning the behavior of electrons was inferred' from the inconsistent theory, i.e. as not every q was inferred from the theory, Bohr must have

used paraconsistent reasoning (Priest 1996). If one combines the insight that some scientific theories are inconsistent with a belief that we are TMs, one must conclude that we are paraconsistent TMs.

The phenomenon of belief revision also suggests that we use paraconsistent reasoning (see Restall and Slaney (1995), and Priest (1996)). We believe something because we think it true, yet, humbly, we do allow for the possibility that at least one of our beliefs could be—and probably is—false; if we didn't allow for this possibility, it's hard to see how belief revision would even be possible (as we would never reject a held belief as false). The so-called 'paradox of the preface' captures this idea somewhat: 'a rational person, after thorough research, writes a book in which they claim $A,\ldots.A_n$,' yet, the author is aware that 'no book of any complexity contains only truths' (Priest 1996), i.e. they also rationally assert not $A,\ldots.A_n$. In short, it appears that we must be paraconsistent to allow for the phenomenon of belief revision, and if one holds that we are paraconsistent, then the paradox of the preface dissolves.

Finally, I point out that I am making a purely descriptive, as opposed to a normative, claim. I am simply claiming that we are in fact paraconsistent, and I leave questions as to whether or not we *should* reason in the manner that we do to the side as irrelevant to my thesis.

Implications for Lucas-Penrose

The claim, or even merely the possibility, that we are paraconsistent has serious implications for L-P; here, I briefly point out two. First, Lucas' attempt to respond to Putnam's (1960, 95) classic criticism of L-P fails. Second, and more seriously, L-P is collapsed into a disjunction, and hence prevented from reaching its desired conclusion.

As is well known, H. Putnam (1960, 95) put forth what is generally recognized as the most serious objection to L-P; S. Guccione (1993, p.62), for instance, calls Putnam's criticism the 'most conclusive and immediate objection' to L-P. Putnam's (1960, 95) criticism can be summarized in the following manner:

(1) Gödel's First Incompleteness Theorem only applies to consistent formal systems.

(2) Gödel's Second Incompleteness Theorem establishes that one cannot establish the consistency of a formal system from within the system itself, so

(3) If we are TMs, we can never establish our own consistency (and cannot, therefore, confidently apply Gödel's First Theorem to ourselves, i.e. L-P is a nonstarter).

Lucas has attempted to meet Putnam's (1960, 95) objection by pointing out that if we were inconsistent, we'd assert any random q (as any q can be inferred in a contradictory system). Since we don't assert any random q, we must be consistent, that is, here we have a consideration that can establish our consistency and thereby overcome Putnam's objection to L-P.

Now, as I pointed out in the previous section, one cannot infer any q in paraconsisent systems. If, as I argued above, we are paraconsistent, then we would have an alternative explanation for why we don't assert any q from the one Lucas puts forth. In effect, Lucas' attempt to overcome Putnam's devastating criticism is inadequate, or at least inconclusive, given the possibility that we are paraconsistent.

But, the possibility that we are paraconsistent raises an even more serious problem for L-P. Recall that in paraconsistent systems, it may very well be possible to prove the G. sentence. If one carries this insight over into the context of the debate surrounding L-P, one sees that a possible explanation for why we can decide the G. sentence is that we are paraconsistent TMs, and not, as L-P argues, that we are not TMs at all. In short, L-P holds that we are not TMs because we can decide the G. sentence while no TM can, but if one allows for the possibility that we are paraconsistent, one sees that the most L-P can conclude from the realization that we can decide the G. sentence is the following disjunction: *either* we can decide the G. sentence because we are not TMs *or* we can decide the G. sentence because we are paraconsistent TMs.

As is well-known, Benacerraf's (1967) famous objection to L-P pursued a similar strategy, i.e. Benacerraf also collapsed L-P

into a disjunction. This suggests that perhaps this novel objection is at least as serious as Benacerraf's (1967) objection.

References

Anderson, A.R. and Belnap, N. D. Jr. *Entailment: The Logic of Relevance and Necessity*. Princeton, Princeton University Press, Vol. I. Anderson, A.R., Belnap, N.D. Jr. and Dunn, J.M. *Entailment*, Vol. II, 1975.

Benacerraf, P. "God, the Devil and Gödel," *The Monist* 51, 9-32, 1967.

Dennett, D.C. "Cognitive Wheels: The Frame Problem in A.I." In *Minds, Machines and Evolution* (Ed: Hookaway). Cambridge: Cambridge University Press, 1984.

Gödel, K. *On Formally Undecidable Propositions in Principia Mathematica and Related Systems*, 1931. (Trans: Meltzer). Edinburgh and London: Oliver Boyd, 1962.

Guccione, S., *Journal of Behavioral and Brain Sciences*, 16:3, p. 612, 1993.

Fodor, J.A. *The Language of Thought*. New York: Crowell, 1975.

Fodor, J.A. *Representations*. Cambridge, MA: MIT Press, 1981.

Fodor, J.A. *Psychosemantics*. Cambridge, MA: MIT Press, 1987.

Fodor, J.A. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press, 1990.

Fodor, J.A. *The Elm and the Expert*. Cambridge, MA: MIT Press, 1993.

Lucas, J.R. "Minds, Machines and Gödel.," *Philosophy*, 36: 112-127, 1961.

Lucas, J.R. *The Gödelian Argument: Turn Over the Page*. Copyright J.R. Lucas. http://users.ox.ac.uk/~jrlucas/turn.html.

Mares, E. "Relevance Logic," *Stanford Encyclopedia of Philosophy*. Copyright E. Mares, 1998.

Penrose, R. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. New York: Oxford University Press, 1989.

Penrose, R. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. New York: Oxford University Press, 1994.

Priest, G. and Tanaka, K. "Paraconsistent Logic," *Stanford Encyclopedia of Philosophy*. Copyright G. Priest and K. Tanaka, 1996; 2000.

Priest, G. Routley, R. and Norman, J. (Eds). *Paraconsistent Logic: Essays on the Paraconsistent*. Munchen: Philosophia Verlag, 1989.

Priest, G. "Paraconsistent Logic," *Handbook of Philosophical Logic* (2nd Edition). Forthcoming.

Putnam, H. "Brains and behavior," 1960, Reprinted in *Readings in the*

Philosophy of Psychology, pp. 24-36. (Ed: N. Block). Cambridge, MA: Harvard University Press, 1980.

Putnam, H. "The nature of mental states," *Art, Mind and Religion* (Eds: W. H. Capitan and D.D. Merrill). Pittsburgh: University of Pittsburgh Press. Reprinted in *Readings in the Philosophy of Psychology* (Ed: N. Block). Cambridge, MA: Harvard University Press, 1967.

Putnam, H. "Review of R. Penrose's Shadows of the Mind: A Search for the Missing Science of Consciousness," *Bulletin of the American Mathematical Society*, Vol. 32, pp. 370-73, 1995.

Pylyshyn, Z. "Computation and cognition: Issues in the foundation of cognitive science," *Behavioral and Brain Sciences* 3: 111-32, 1980.

Pylyshyn, Z. *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, MA: MIT Press, 1984.

Restall, G. and Slaney, J. (1995). *Realistic Belief Revision, Technical Report: TR-ARP-2-95*, Automated Reasoning Project, Australian National University, 1995.