Why We Shouldn't Reason Classically, and the Implications for Artificial Intelligence

Douglas Campbell

University of Canterbury
Christchurch, New Zealand
douglas.campbell@canterbury.ac.nz
+64 3 3642987 x6859

In this paper I argue that human beings should reason, not in accordance with classical logic, but in accordance with a weaker 'reticent logic'. I characterize reticent logic, and then show that arguments for the existence of fundamental Gödelian limitations on artificial intelligence are undermined by the idea that we should reason reticently, not classically.

1. Introduction

In this paper I argue that human beings should reason, not in accordance with classical logic (CL), but in accordance with what I will call 'reticent logic' (RL). To see why we shouldn't reason classically, imagine two prisoners, locked in Cells A and B respectively. Each prisoner is given a list of sentences, and can choose whether to 'accept' sentences in the list. We can suppose a prisoner accepts a sentence by checking a box next to it. I am one of the prisoners. Initially I don't know whether I am in Cell A or Cell B, but I know it will be announced soon which cell I am in.

My list looks like this:

1: If I am Cell B's inmate, then Cell B's inmate will never accept 3. ☐
2: I am Cell B's inmate. ☐
3: Cell B's inmate will never accept 3. ☐

My aim is to accept only true sentences. (E.g., imagine one year will be deducted from my prison-sentence for each true sentence I accept, but a year added for each false sentence.) If Cell B's inmate was to accept 3, then 3 would be false, and so Cell B's inmate would have accepted a falsehood. Recognizing this, I resolve that if it is announced that I am Cell B's inmate I will never accept 3.[1] My track record of following through on such resolutions is perfect. Hence I have good grounds for thinking that if I am Cell B's inmate, then Cell B's inmate will never accept 3. This is what 1 says. Accordingly I accept 1, by checking its box.

Next it is announced that I am Cell B's inmate, and so I check 2's box.

CL includes the rule of inference, *modus ponens*, which validates the inference from 1 and 2 to 3. Thus $1,2 \vdash 3$ (henceforth, the *prisoner's argument*) is classically

[1] If the word 'never' raises intuitionistic worries about permanently undetermined truth-values, then it

valid. So were I to reason classically then I would, having accepted 1 and 2, also accept 3. However, I would falsify both 1 and 3 by accepting 3.[2]

What are my options? There appear to be four:

(i) I might accept 1 and 2, reason classically, and accept 3. This is a bad option, for as just seen it results in me accepting only one truth and two falsehoods.

(ii) I might accept 1 and 2, but refuse to accept 3, even though 3 is classically entailed by 1 and 2. This option is attractive, since it results in me accepting two truths and no falsehoods. However it means I must reason non-classically.

(iii) Foreseeing a trap, I might refuse to accept either 1 or 2 so CL won't push me into accepting 3. This option is unattractive for two reasons. First it results in me accepting only one truth, instead of the two truths I get to accept under option (ii). (If I am to refuse to accept a sentence I can plainly see to be true, better it be 3 rather than 1 or 2.) Second, both 1 and 2 might be classically entailed by other statements I can see to be true, creating a risk of escalation: to avoid being forced by CL into accepting 3, I might have to refuse to accept, not only 1 or 2, but numerous other true propositions from which 1 and 2 can be derived.

(iv) 3 is self-referential, and in this respect similar to the strengthened liar sentence ('This sentence is untrue'), which lacks coherent truth-conditions. It might be suggested on this basis that 3 lacks coherent truth-conditions too. If this were right then 3 wouldn't be classically entailed by 1 and 2, dissolving the problem. However, this option appears untenable. The strengthened liar sentence is paradoxical because any attempt to assign it a truth-value yields contradiction: the supposition it is true supports the conclusion it is untrue, and *vice versa*. In contrast, neither the supposition that 3 is true nor the supposition it is untrue is contradictory. Rather, to suppose 3 is true is merely to suppose that Cell B's inmate never checks the third box on his list, while to suppose 3 is untrue is to suppose that Cell B's inmate will eventually check this box. 3's truth conditions are therefore unproblematic.[3]

Since (ii) is the best of these options, the prisoner's argument provides strong *prime facie* support for the idea that we shouldn't reason classically. But according to which logic should we reason, if not CL? This paper is structured as follows. §2 introduces the key notion of a 'perverse argument'. §3 describes RL and argues we should reason reticently, rather than classically. It also briefly outlines several sub-varieties of RL. §4 and §5 showcase philosophical applications of the claim that we should reason reticently, with §4 critiquing a Gödellian argument against the possibility of an artificial intelligent machine knowing itself to be consistent, and §5 critiquing the famous 'mathematical argument' against artificial intelligence. §6 wraps things up.

2. Perverse arguments

To 'accept' a sentence is to perform some mechanical action by which one endorses it as being true. For example, in the scenario just discussed the prisoner 'accepts' a sentence by ticking a box next to it. A formal system can be regarded as

[2] Is the prisoner's argument a counterexample to *modus ponens*? No—or at least, not if by 'counterexample' we mean a case where both ϕ and $\phi \rightarrow \psi$ are true but ψ is false. The prisoner's argument is instead a case in which ϕ and $\phi \rightarrow \psi$ can both be true only if ψ is not accepted.

[3] Lingering suspicions that 3 is liar-like should be put to rest by noticing that Gödel's (1931) diagonalization procedure for generating self-referential sentences with well defined truth-conditions can be used to manufacture a version of 3. See §5, below, for an explanation of how this procedure can be applied to English.

'accepting' a sentence by proving it as a theorem. A person can be regarded as 'accepting' a sentence, ϕ, by believing ϕ (i.e., by loading ϕ into her 'belief box', as it were), or by saying, "ϕ is true" or "I accept ϕ". The notion of acceptance is intended to be a general one, having each of these other notions as special cases.

Notation. Let $\Box\phi$ be shorthand for 'This system will ultimately accept ϕ' (or 'I will ultimately accept ϕ').[4] So, if a system accepts both ϕ and $\Box\phi$, then it thereby ensures that the latter sentence is true by accepting the former sentence. On the other hand, if it accepts $\Box\phi$ but never accepts ϕ, then in accepting $\Box\phi$ it accepts a falsehood.

With this notation in place, the prisoner's argument is revealed as having the following form:

A0. $(P \wedge Q) \rightarrow \neg\Box Q$
A1. $P \rightarrow Q$
A2. P

A3. Q

Here P stands for 'This system is Cell B's inmate'. Q stands for 'Cell B's inmate will never accept Q'. A0 isn't an explicit premise of the prisoner's argument, but is a tautological adjunct to the argument. It says, 'If this system is Cell B's inmate and Cell B's inmate will never accept Q, then it is not the case that this system will ultimately accept Q'.

A0, A1 and A2 together classically entail both Q and $\neg\Box Q$. That is, they classically entail both that Q is the case and that Q won't be accepted by the system. Let such arguments be called *perverse*. I.e., an argument is perverse iff: (a) its conclusion, ϕ, is classically entailed by its premises (i.e., the argument is classically valid); and (b) $\neg\Box\phi$ is also classically entailed by its premises. More generally, a proposition-set is perverse iff there is some ϕ such that the set classically entails both ϕ and $\neg\Box\phi$.

Perversity isn't to be confused with inconsistency. For example, the prisoner's argument's premises are perverse and yet clearly consistent (as can be seen by noticing that if I am Cell B's inmate and I never accept 3, then both 1 and 2 will be true).

Let $S \vdash \phi$ be some perverse argument. A system which reasons classically from S will commit a kind of fallacy—the 'perversity fallacy' as I shall call it. In accepting S, it is committed, on pain of having accepted a falsehood, to not accepting ϕ (since S entails $\neg\Box\phi$), and yet because it reasons classically and S classically entails ϕ, it *will* accept ϕ. Thus by accepting ϕ it ensures the falsity of S, *thereby undermining its grounds for concluding that ϕ is true in the very act of drawing this selfsame conclusion*. Such a classical reasoning system is like a moth flying in the dark near a candle. Just as the moth's method of navigation dooms it to the flame, so a system that reasons classically will blunder inevitably into error if a perversity lurks in the base of sentences it is reasoning from.

[4] I borrow the '$\Box$' notation from provability logic, wherein the intended meaning of '$\Box\phi$' is 'ϕ is provable in Peano Arithmetic'. In using this notion I don't mean to suggest that RL is a standard modal logic. (It isn't.)

3. Reticent Logic (RL)

The idea behind RL is that to avoid succumbing to the perversity fallacy we should always do a 'perversity check' before accepting the conclusion of a classically valid argument. A reticent logic (RL) is simply a logic that includes a perversity check. Such a logic is 'reticent' in the sense that it 'holds back' in some cases when CL blithely accepts the conclusion of a perverse argument.

This idea can be implemented in various ways, of varying sophistication, some of which are now briefly explained.

3.1 Basic RL

Basic RL classifies arguments as *reticently valid* or *reticently invalid*. $S \vdash \phi$ will be classified as reticently valid if these two conditions are satisfied:

(a) $S \vdash \phi$ is classically valid.
(b) $S \vdash \neg\Box\phi$ is not classically valid (i.e., $S \vdash \phi$ passes the 'perversity check').[5]

Otherwise $S \vdash \phi$ is classified as reticently invalid.

For example, although the prisoner's argument is classically valid, Basic RL classifies it as reticently invalid. This is because A0, A1 and A2 classically entail not only Q, but also $\neg\Box Q$.

Basic RL is weaker than CL, in the sense that while every reticently valid argument is classically valid, some classically valid arguments are not reticently valid. It can be thought of as being a logic of two parts, these being: (i) CL's methods for classifying an argument as classically valid or classically invalid; and (ii) a 'devalidating rule' that reclassifies perverse classically valid arguments as 'invalid'. In other words, it is a logic that sets the bar for validity higher than CL, by demanding not only that it be impossible for the premises to be true whilst the conclusion is false, but also that it be possible for the premises to be true whilst the conclusion is accepted.

3.2 Stepwise RL

By a 'logic' we usually mean not just a method for classifying arguments as valid or invalid, but a set of rules of inference that allow the conclusion of a valid argument to be derived from the argument's premises through a series of intermediate steps. Perversities might lurk at any step. A *stepwise RL* is a version of RL that performs a perversity check at each step. It consists of a set of *reticent rules of inference*, that differ from the classical rules by dint of having perversity checks built into them. For example, the classical and reticent versions of *modus ponens* differ from each other as follows:

---

[5] The perversity check will be straightforward if the language is that of propositional logic or unary predicate logic, since it will then be decidable whether $S \vdash \neg\Box\phi$ is classically valid. For richer languages it will be necessary to make do with an incomplete perversity testing method, that errs by sometimes failing to classify perverse arguments as perverse. For every such method there will be a corresponding version of Basic RL, with its own strengths and weaknesses where its ability to detect perversities is concerned. The question as to which of such methods are 'best' is rich and complex, but I say no more about it here.

Classical modus ponens: if both ψ and $\psi \rightarrow \phi$ are accepted, then accept ϕ.

Reticent modus ponens: if, (i) both ψ and $\psi \rightarrow \phi$ are accepted, and (ii) $\neg\Box\phi$ isn't classically derivable from any sentences that are already accepted, then accept ϕ.

Detecting whether condition (ii) is satisfied requires a meta-level test to be conducted, to see whether $\neg\Box\phi$ is classically entailed by the sentences accepted to date. Doing this meta-level test for classical validity will typically require invoking the ordinary, classical rules of inference multiple times. Hence the reticent rules of inference presuppose the classical rules. One can therefore accept this paper's thesis—that we should reason reticently rather than classically—while still maintaining that there remains a strong sense in which CL is the most fundamental logic.

3.3 Weak RL

Consider Argument B:

B1. P
B2. $P \rightarrow Q$
B3. $(\Box P) \rightarrow \neg\Box Q$

B4. Q

Is this argument perverse? No: for the premises classically entail only Q, not $\neg\Box Q$. But there is a similar fallacy involved in accepting the conclusion of this argument if one accepts all its premises. In accepting B1 one ensures that $\Box P$ is true. $\Box P$ and B3 together classically entail $\neg\Box Q$: i.e., that Q will not be accepted. Thus by accepting Q, one would falsify B3 and make it the case that one has accepted a falsehood.

Next, consider Argument C:

C1. $\Box P$
C2. $(\Box P) \rightarrow Q$
C3. $P \rightarrow \neg\Box Q$

C4. Q

Again this argument isn't perverse but a similar fallacy lurks in it. By accepting C1, one commits oneself, on pain of having accepted a falsehood, to accepting P. In accepting P together with C3 one is committed to the truth of $\neg\Box Q$—i.e., to not accepting the argument's conclusion. So if one accepts the conclusion in addition to the premises then one has accepted at least one falsehood.

Finally, consider Argument D:

D1. $\neg P$
D2. $\neg P \rightarrow Q$
D3. $\neg\Box P \rightarrow \neg\Box Q$

D4. Q

Yet again, this argument is not perverse but contains a similar fallacy. By accepting D1 (which is to say, $\neg P$), one is committed, on pain of contradicting oneself, to not accepting P. That is, one is committed to the truth of $\neg\Box P$. But $\neg\Box P$ and D3 together classically entail $\neg\Box Q$. So by accepting D1, D2 and D3 one in effect commits oneself, on pain of having accepted a falsehood, to not accepting the argument's conclusion, Q.

To enable the detection and avoidance of fallacies like the above, a rule of inference, U, and axiom, V, may be introduced.

Rule U: If any formula, ϕ, is accepted, then $\Box\phi$ may be accepted too.

Axiom V: $(\Box\phi)\rightarrow\phi$

Rule U is obviously well motivated. It enables a system that has accepted ϕ to accept it has done so—i.e., to accept $\Box\phi$. (It makes the system 'self conscious', so to speak.) It is 'truth preserving', since it will never directly cause a system to accept a falsehood.

Axiom V is similarly well motivated, for upon accepting $\Box\phi$, a system is committed, on pain of having accepted a falsehood, to accepting ϕ too. V lets the system discharge this commitment. The inference step from $\Box\phi$ to ϕ is truth preserving in the sense that a system that has accepted $\Box\phi$ has 'burnt its bridges' and can only hope to keep its set of accepted sentences free of falsehoods by accepting ϕ too.[6]

V is logically equivalent to its contrapositive, V′:

V′: $\neg\phi\rightarrow(\neg\Box\phi)$

V′ allows a system to derive $\neg\Box\phi$ from $\neg\phi$. Again this inference step appears reasonable: for if a system has accepted $\neg\phi$ then it is committed, on pain of contradicting itself, to not accepting ϕ, and thus to the truth of $\neg\Box\phi$.[7]

Let *UV logic* be a logic obtained by adding U and V (and hence V′) to CL. An argument's premises *UV-entail* its conclusion iff the conclusion is derivable from the premises using UV logic. An argument is *neo-perverse* if it is not perverse, but if it is such that: (i) its premises classically entail its conclusion, ϕ; and (ii) its premises UV-entail $\neg\Box\phi$.

---

[6] From the fact that the system has accepted $\Box\phi$, it does not follow that ϕ is *true*. But it does follow that if the system fails to accept ϕ, then its risk of having accepted a falsehood is 100%.

[7] Suppose a set of premises are inconsistent, classically entailing both ϕ and $\neg\phi$. Suppose too that $\neg\phi$ is accepted. V′ then permits $\neg\Box\phi$ to be accepted, which will then cause RL's devalidating rule to prevent ϕ being accepted. Thus the system will be 'consistent' from the perspective of RL (in the sense that both ϕ and $\neg\phi$ will not be derivable) despite being classically inconsistent. Obviously, this variety of guaranteed consistency is trivial and uninteresting. Should we reject V′ (and thus V) because it 'debases' consistency in this way? I think not. After all, CL also yields a peculiar result – 'logical explosion' – when premises are classically inconsistent. RL is designed to keep us out of trouble, not when our premises are classically inconsistent (in which case a paraconsistent logic is called for), but when our premises are classically consistent but contain a perversity or neo-perversity. V′ is vital in enabling certain neo-perversities to be detected. (Many thanks to Doukas Kapantais for comments on this point.)

For example, Arguments B, C and D are neo-perverse. They are neo-perverse because their premises classically entail their (shared) conclusion, Q, and because their premises also UV-entail $\neg\Box Q$. The three arguments differ with respect to whether it is U, V or V′ that is crucial in enabling the neo-perversity to be detected. With Argument B, U is crucial (allowing $\Box P$ to be derived from B1). With Argument C, it is V that is important (allowing P to be derived from C1). And with Argument D, V′ plays the key role (allowing $\neg\Box P$ to be derived from D1).

Let the *neo-perversity fallacy* be the fallacy of accepting the conclusion of a neo-perverse argument. For instance, one would commit this fallacy by reasoning classically from the premises of Arguments B, C or D. Let a *Weak Reticent Logic* (Weak RL) be a version of RL that includes not only a perversity check, but also a neo-perversity check, and which is therefore capable of detecting and avoiding both perversity fallacies and neo-perversity fallacies. To be more precise, it is a version of RL such that the conclusion, ϕ, of a classically valid argument will be accepted only if $\neg\Box\phi$ isn't UV-entailed (or classically entailed) by the argument's premises.

As explained above, Basic RL is weaker than CL, in the sense that it validates only a proper subset of the arguments validated by CL. 'Weak RL' is so-called because it validates only a proper subset of the argument validated by Basic RL, making it an even weaker logic than Basic RL.

An explanation is in order as to why U is a *rule of inference*, instead of being an *axiom*, like V. What would be wrong with replacing U with the axiom, W?[8]

W: $\phi\rightarrow(\Box\phi)$

It might appear that W achieves the same effect as U. For instance, W would, like U, enable us to detect that Argument B is neo-perverse, by allowing us to derive $\Box P$ from B1.

To see the problem, notice that W is logically equivalent to its contrapositive, W′:

W′: $(\neg\Box\phi)\rightarrow\neg\phi$

W′ is problematic. Suppose one has accepted a perverse set of sentences that classically entail both ϕ and $\neg\Box\phi$. RL tells one not to accept ϕ in this situation. But of course this is not to suggest that one should accept ϕ's denial, $\neg\phi$. To the contrary, one obviously shouldn't accept $\neg\phi$ in such cases, for were one to accept $\neg\phi$ in addition to the sentences one already accepts (which entail ϕ), then one would be inconsistent. W′ would allow one to derive $\neg\phi$ from $\neg\Box\phi$ in such cases. For this reason W′ can't have any place among the axioms of RL, and nor can W, which is logically equivalent to W′.[9]

[8] It is immediately clear that W and V cannot *both* be axioms: for together they entail $\phi\leftrightarrow(\Box\phi)$ and so drain the '□' operator of significance (Jack Copeland, personal communication).

[9] Why can't W (and thus W′) be derived using U, as follows? First, assume ϕ. Then use U to derive $\Box\phi$ from ϕ. Then, by the rule of conditional introduction, derive $\phi\rightarrow\Box(\phi)$. Answer: in RL it is crucial to distinguish between propositions that have the status of being *accepted* and propositions that merely have the status of being *assumed* for the purpose of proving a conditional. U can only be used to derive $\Box\phi$ from ϕ when ϕ has been accepted, not when ϕ has only been assumed.

3.4 Strong RL

The above versions of RL are all strictly weaker than CL, in the sense that they each validate only a proper subset of the inferences validated by CL (those that are not perverse and/or neo-perverse). What I call a 'Strong RL' is, on the other hand, a version of RL that validates certain inferences that are not classically valid. It does this by using U and V, not only at the meta-level where inferences are checked for perversities and neo-perversities, but also at the base level where inferences are first generated. That is, it differs from Weak RL in that, whereas Weak RL uses CL at the base level and UV-logic at the meta-level, Strong RL uses UV-logic at both levels.

For example, consider Argument E:

E1. P
E2. $\Box Q$
E3. $(\Box P \wedge Q \wedge \neg\Box\neg P) \rightarrow R$
E4. R

This argument isn't classically valid, and so, of course, it isn't valid by the lights of Weak RL, either. But someone who has accepted its premises is nonetheless committed to the truth of its conclusion. To see this, notice that in accepting E1, one makes it the case that $\Box P$ is true. Similarly, in accepting E2 one commits oneself, on pain of having accepted a falsehood, to accepting Q, and thus one commits oneself to Q's truth. Finally, in accepting E1, one commits oneself, on pain of contradicting oneself, to not accepting $\neg P$ (i.e., to the truth of $\neg\Box\neg P$). Hence in accepting both E1 and E2 one is committed to the truth of every conjunct of E3's antecedent. Thus if one accepts E3 too, one is committed to the truth of E4. In short, Argument E is, if not classically valid, still 'valid' in some clear sense of that term. It is precisely this species of validity that is analyzed by Strong RL. Strong RL classifies Argument E as valid, because E4 can be derived from E1, E2 and E3 with the help of U and V. (Specifically, U enables $\Box P$ to be derived from E1; V enables Q to be derived from E2; and V's contrapositive, V′, enables $\neg\Box\neg P$ to be derived from E1.)

3.5 Classifying RL

RL is a non-monotonic logic, since adding $\neg\Box\phi$ to premises that reticently entail ϕ yields premises that don't reticently entail ϕ. It is a deductively incomplete logic since there can be a ϕ such that neither ϕ nor $\neg\phi$ is reticently entailed by the premises (as when the premises classically entail both ϕ and $\neg\Box\phi$).

RL has some resemblance to a modal logic. For example, RL's U amounts to a strengthened version of K's Necessitation Rule, and RL's V is identical to modal logic's axiom M. However, RL doesn't respect K's Distribution Axiom, $\Box(\phi\rightarrow\psi)\rightarrow(\Box\phi\rightarrow\Box\psi)$, and so it is certainly not a standard modal logic.

4. Can a consistent artificially intelligent machine prove it is consistent?

The project of developing RL in detail is a book-length one. I leave it for another occasion. The remainder of this paper is devoted instead to showing that RL has important philosophical applications.

Consider Argument G:

G1. F is consistent

G2. If *F* is consistent, then *F* will not prove *G(F)*

G(F). *F* will not prove *G(F)*

Here *F* denotes some formal system, and *G(F)* is an English rendering of *F*'s Gödel sentence. Gödel (1931) showed that, provided *F* uses various classical rules of inference such as *modus ponens* (see, e.g., Raatikainen, 2014) and encodes elementary arithmetic, then if *F* proves its own consistency (i.e., if it proves G1) it will be driven, by Argument G, to accept *G(F)*, from which it follows that *F* is actually inconsistent. Since digital computers amount to instantiations of formal systems, this result – Gödel's second incompleteness theorem – can be taken (see, e.g., Gaifman, 2000) as implying that no artificially intelligent digital computer can prove its own consistency except on pain of inconsistency.

I believe that Gödel's second incompleteness theorem has no such implications. To see why not, let us analyze 'proof' as being a species of 'acceptance', and use '$\Box\phi$' to represent the claim, '*F* will prove ϕ'. Argument G can then be formalized as follows:

G1′. *Con(F)*

G2′. $Con(F) \rightarrow \neg\Box G'(F)$

$G'(F)$. $\neg\Box G'(F)$

This argument is perverse, since its premises entail both $G'(F)$ and (same thing) $\neg\Box G'(F)$. Needless to say this perversity is crucial to Gödel's argument, since his strategy for proving the second incompleteness theorem hinges entirely on the idea that *F* will falsify Argument G's conclusion, and thereby falsify one of the argument's premises (namely, G1) in the very act of proving this selfsame conclusion.

Let us suppose that *F* is an artificially intelligent system that reasons reticently, rather than classically. In this case Gödel's second incompleteness theorem will not apply to it. While the theorem applies to any formal system that satisfies certain, modest requirements, one of these requirements is that the system uses classical logic. If *F* does not obey the various rules of classical logic, including *modus ponens*, then *F* need not be driven by the logic it is using from proving G1 and G2 to proving *G(F)*. Indeed if – as we are imagining – the non-classical logic used by *F* is RL, then *F* certainly *will not* prove *G(F)* after accepting G1 and G2: for it will instead recognize that Argument G is perverse and refuse to prove *G(F)* for this reason. Because it won't prove *G(F)*, it won't be caused, by its having proved G1 and G2, to undermine its own consistency by proving *G(F)*. Thus – at least for all Gödel's argument shows – it is entirely possible for such a system, which reasons reticently rather than classically, to prove both G1 and G2 (and thus prove its own consistency) without thereby tumbling into inconsistency.

For the reasons just given, it appears that consistent artificially intelligent computers will be unable to prove their own consistency only if they must reason classically, rather than reticently. But why couldn't an artificially intelligent machine reason reticently? Why not indeed! It is surely plausible that any machine that is truly 'intelligent' will be capable of recognizing whether, in accepting various premises, it has committed itself to not proving a conclusion that follows classically from these premises, and of refusing to prove the conclusion in such cases. That is, machines that are genuinely intelligent will surely not be prone to succumbing to fallacies of perversity and neo-perversity. They will use RL, not CL.

5. A rebuttal of the mathematical argument against artificial intelligence

The 'mathematical argument' against artificial intelligence (Gödel, 1951; Nagel and Newman, 1957; Lucus 1961, 1996; Penrose 1989, 1994, 1996) purportedly shows that the theorem-proving abilities of the human mind cannot be matched by a computer. There is widespread agreement among philosophers and mathematicians that the argument is defective, but less agreement as to why. In what follows I first consider several stock rejoinders to the argument, and show that the argument can be patched to avoid them. Next I contend that the real problem with the argument involves a perversity fallacy within it.

Let the mathematical argument's 'protagonist' – referred to in the first person – be some human mathematician. Let *F* be some formal system (or programmed digital computer). 'I am *F*' is the conjecture that the protagonist's sentence proving dispositions match *F*'s. Let this conjecture be called the 'identity hypothesis'. The original version of the mathematical argument, found in Lucus (1961) and Penrose (1989), may be summarized as follows:

H1.	I am consistent (i.e., I won't, for any sentence ϕ, prove both ϕ and $\neg\phi$).
H2.	If I am *F*, then *F* is consistent. (From H1.)
H3.	If *F* is consistent, then I can prove *F* is consistent.
H4.	If I can prove *F* is consistent then (by invoking Gödel's first incompleteness theorem) I can know that *G(F)* is true.
H5.	If I can know that *G(F)* is true, then I can prove *G(F)* without compromising my consistency.
H6.	If I am *F*, then I can prove *G(F)* without compromising my consistency. (From H2 – H5.)
H7.	If I am *F*, then I cannot prove *G(F)* without compromising my consistency. (From Gödel's first incompleteness theorem.)
H8.	I am not *F*. (From H6 and H7, by *reductio*.)

The most glaring point of weakness in Argument H is H3. Many authors (e.g., Putnam, 1960; Bowie, 1982; Barr, 1990; Boolos, 1990; and Gaifman, 2000) have pointed out that there is ample room to imagine that: (i) the protagonist's sentence-producing powers might be equivalent to those of some consistent formal system, *F*; but that (ii) the protagonist might be unable, because of *F*'s great complexity, to prove that *F* is consistent.

Penrose has developed an ingenious new version of the mathematical argument that sidesteps this problem (1994, pp. 179-188; 1996). It is sometimes called Penrose's 'new argument', but I will call my formulation of it 'Argument J'. [10] It is based on the idea that we need not require the protagonist to prove that *F* is consistent 'from the ground up', as it were, because we can instead start from the assumption (contained in H1) that the protagonist herself is consistent, and then

[10] There is some question as to precisely how Penrose's 'new argument' is supposed to go (see Chalmers 1995; Penrose 1996; Lindström 2001, 2006; Shapiro 2003). My Argument B is closely based on Penrose's (1994) original, informal presentation of the argument, and its essential logic is similar to Lindström's (2001) formulation. Departing from Penrose, I frame the argument in terms of the *consistency* of the formal systems in question, instead of the *soundness* of these systems, with the reason being that the former notion is less demanding and more general than the latter but still adequate for the argument's purposes.

cantilever sideways from this starting point to the conclusion that, if the identity hypothesis is correct, then *F* must be consistent too. The argument requires as a premise not only that the protagonist is consistent, but also that she *knows* she is consistent. Since she knows she is consistent, she can know that, if 'I am *F*' is true, then *F* is consistent. She doesn't know whether *F* is *in fact* consistent (because she doesn't know whether 'I am *F*' is true), but as an intellectual exercise she can *imagine* that 'I am *F*' is true and explore the logical consequences of this supposition. In doing this she will prove various sentences of the form, *if I am F then* ϕ. If 'I am *F*' is *in fact* true then any such sentence that she can prove will also be proved by *F*. Penrose has us consider another formal system *F′*, which is like *F* but which internalizes 'I am *F*' as an extra axiom. Thus, if *F* proves any sentence of the form *if I am F then* ϕ, *F′* will instead simply prove ϕ. Penrose observes that if F is consistent, and if 'I am *F*' is true, then *F′* must be consistent too. Thus the identity hypothesis implies, not only that *F* is consistent, but also that *F′* is consistent, and thus (via Gödel's theorems) that *F*'s Gödel sentence, *G(F′)*, is true. The argument's protagonist can recognize this (for *we* can recognize this), so she can prove the sentence, 'if I am *F*, then *G(F′)*'. If the identity hypothesis is in fact correct, then *F* will prove this sentence too. But if *F* proves this sentence, then *F′* will prove *G(F′)*, which, by Gödel's theorem, is something it cannot do if it is consistent. In short, the identity hypothesis implies both that *F′* is consistent, and that *F′* will prove *G(F′)* – contradicting Gödel's theorem. Hence the identity hypothesis must be false. More formally:

J1.	I am consistent, and I know it.
J2.	If I am *F*, then *F* is consistent. (From J1.)
J3.	If I am *F* and *F* is consistent, then *F′* is consistent (where *F′* is a formal system obtained by adding an extra axiom, 'I am *F*', to *F*, so that *F′* proves ϕ iff F proves 'If I am *F* then ϕ').
J4.	If I am *F*, then *F′* is consistent. (From J2 and J3.)
J5.	If *F′* is consistent, then *G(F′)*. (From Gödel's first incompleteness theorem.)
J6.	If I am *F*, then *G(F′)*. (From J4 and J5.)
J7.	I know that J1, J3, and J5 are true.
J8.	If I know that J1, J3 and J5 are true, then I know that I can, by proving J6, prove a truth (since I can see that J6 follows from J1, J3 and J5).
J9.	If I know that I can, by proving J6, prove a truth, then I will prove J6.
J10.	I will prove J6. (From J7 – J9.)
J11.	If I am *F*, and if I will prove J6, then *F* will prove J6.
J12.	If *F* will prove J6 then *F′* will prove *G(F′)* (since J6 is of the form 'if I am *F*, then ϕ', with *G(F′)* replacing ϕ). (From what J3 says about *F′*.)
J13.	If I am *F*, then *F′* will prove *G(F′)*. (From J11 and J12.)
J14.	If *F′* is consistent, then *F′* will not prove *G(F′)*. (From Gödel's first incompleteness theorem.)
J15.	If I am *F*, then *F′* will not prove *G(F′)*. (From J4 and J14)
J16.	I am not *F*. (From J13 and J15, by *reductio*.)

Three objections to Argument J are now considered.[11] The most popular objection targets the claim that the protagonist is consistent and knows it (i.e., premises J1). For instance, according to Turing (1947, 1948, 1950) the moral to be drawn from Gödel's work is that one can be intelligent enough to reason about the incompleteness theorems only if one is also so prone to error that no confidence can be put in the consistency of one's beliefs. In Turing's words, 'if a machine is expected to be infallible, it cannot also be intelligent' (1947). For Turing, fallibility – and a concomitant ability to make mistakes and then learn from them – is an *essential ingredient* of intelligence. Other authors (e.g., Grush and Churchland 1995) take the less radical position that, even if fallibility is perhaps not *necessary* for intelligence, it is nevertheless such an ineluctable feature of human performance that no human mathematician can know she is consistent.

Argument J can, I believe, be patched up to make it invulnerable to such objections by supposing that the argument's protagonist is what I will call a 'careful typist'. A careful typist is a person who evinces ordinary human fallibility in her day-to-day affairs (and who often makes mistakes and learns from them, as Turing says an intelligent being must), but who is charged with using a typewriter to produce a sequence of true sentences, and who takes the utmost care never to type a sentence unless she has a proof of its truth that meets the most exacting standards of simplicity, rigor and clarity. Whenever in doubt about the truth of a sentence, she errs on the side of caution and doesn't type it. Argument J is silent on what the protagonist must do to 'prove' a sentence. There is therefore nothing to prevent us stipulating that the protagonist 'proves' a sentence by typing it with the typewriter in question. She will therefore be 'consistent' iff the list of sentences she types is free of contradictions. (Mistakes she makes elsewhere in life will be irrelevant.) Provided the protagonist is such a 'careful typist', it is surely plausible that she might both be consistent and know she is consistent.

A similar objection (Chalmers 1995 and McCullough 1995) challenges the claim that the protagonist can know she is consistent (i.e., premise J1) based on Gödellian considerations.[12] Specifically, according to this objection the protagonist will, if the identity hypothesis is true, lapse into inconsistency in the very act of proving herself consistent (i.e., in the act of proving J1). However, as was explained in §4, Gödel's demonstration that a formal system will lapse into inconsistency if it proves itself consistent rests, in part, on the assumption that the system reasons classically, rather than reticently. The present objection is therefore dispensed with by supposing that the protagonist reasons reticently.

The last objection I consider (e.g., Robinson 1992, and Benacerraf 1968) targets premise J7 on the basis that, due to *F″*s complexity, its Gödel sentence, *G(F′)* would be such a stupendously large sentence of arithmetic that the protagonist would be unable to construct it, leaving her unable to know that J5 is true (and thus in no position to prove J6). This objection can be fended off by arranging for the protagonist to use a language in which a syntactically concise version of *F″*s Gödel sentence can be constructed. The following stipulations achieve this result:

- We use some name – say, '$\mathcal{F}$'– as a name for *F′*.

[11] Most of these objections were initially conceived as objections to the original version of the mathematical argument, but apply equally against Argument J.

[12] Chalmers uses Löb's theorem, rather than Gödel's theorem (but these two theorems are intimately related).

- We adopt some arbitrary method (say, some lexicographic method) for assigning Gödel numbers to sentences of English.
- We let *Sub(x,y)* be the Gödel number of the sentence obtained by putting the number *x* in place of each occurrence of the lone free variable (if any) in the sentence with the Gödel number, *y*.
- We let *D(y)* be the diagonalizing sentence, '$\mathcal{F}$ does not prove *Sub(y,y)*'
- We let *d* be *D(y)*'s Gödel number.
- Thus *D(d)* says, '$\mathcal{F}$ does not prove *D(d)*'.[13]

Notice that *D(d)* is a self-referential sentence, that is true iff *F′* does not prove *D(d)*. Thus *D(d)* is – just like G(*F′*) – a Gödel sentence of *F′*. This will come as no surprise since the above 'recipe' for constructing *D(d)* closely mirrors Gödel's own recipe for constructing *G(F′)*, with the only differences being that it uses English instead of Peano Arithmetic and uses the name, $\mathcal{F}$, instead of *F*'s (immensely large) Gödel number.[14] Importantly, whereas the task of constructing *G(F′)* is perhaps beyond the powers of a human, there seems nothing to prevent the protagonist from constructing *D(d)*. All she must do, when presented with a system, *F*, that allegedly models her own mathematical competency, is conceive of *F′* (a system obtained by adding the extra axiom, 'If I am *F*', to *F*), invent a name for it, and use this name in the above recipe. Having constructed *D(d)* in this way, she can use it as a 'stand in' for *G(F′)* within Argument J, as she sets about using this argument to refute the identity hypothesis.

At this point I hope the mathematical argument is beginning to look rather more compelling than it is generally given credit for. For reasons just outlined premises J1 and J7 seem robust. The remaining premises all appear unassailable, being in most cases either tautologies or provable theorems.

So, should we accept the mathematical argument's conclusion, and the implication that human theorem-proving powers exceed those of any formal system or digital computer? I think not. We should instead reject J8:[15]

J8. If I know that J1, J3 and J5 are true, then I know that I can, by proving J6, prove a truth (since I can see that J6 follows from J1, J3 and J5).

J8 appears innocuous at first blush: if one can see that the premises of a manifestly classically valid argument are true, then – so it would seem – one can prove the conclusion, safe in the knowledge that one is proving a truth. But the main theme of this paper has been that such reasoning can be dangerous. We have seen that if an argument is perverse then accepting its premises involves committing oneself, on pain of having accepted a falsehood, to not accepting its conclusion. To prove the conclusion in such a case would be to falsify at least one of premises and commit the perversity fallacy. If the argument, J1, J3, J5 ⊢ J6 is perverse, then J8 is false.

[13] This formulation of Gödel's diagonalization procedure is based on (Rucker 1982, p. 284).
[14] When we ask whether the human protagonist in the mathematical argument can prove things a formal system cannot, we should not force her to use Gödel numbers and Peano arithmetic, which play to the strengths of formal systems, instead of names and natural language, which play to the strengths of the human mind. To do so would be to make her fight with one arm tied behind her back.
[15] The corresponding premise in Argument H is H5, which is problematic for the same reasons as J8.

And, indeed, J1, J3, J5 ⊢ J6 is perverse. To see why it is perverse, we must understand why the premises J1, J3 and J5 together entail, not only that J6 is true, but also that the protagonist will not prove J6. The explanation is as follows. In accepting J1, the protagonist accepts that *she knows she is consistent*. Were she to carelessly prove a sentence that might, as far as she knows, be contradictory, she would not know she was consistent, and so J1 would be false. Hence in accepting J1 she is committed to being careful not to undermine her own consistency. Now, for all that has been shown at this early point in Argument J, the identity hypothesis might be true: i.e., the protagonist's sentence proving dispositions might be the same as *F*'s. So, as part of guarding against undermining her own consistency, the protagonist must be careful not to do anything that would undermine her consistency *if the identity hypothesis happened to be true*. With this thought in mind, let us imagine that the identity hypothesis is in fact true and that the protagonist proves J6. This being so, *F* will prove J6 too. If *F* proves J6, then *F′* proves *G(F′)*. But if *F′* proves *G(F′)*, then, by Gödel's theorem, *F′* is inconsistent. It would follow from this that J4 was false (since J4 says 'if I am *F*, then *F′* is consistent'). But if J4 is false then J1 must be false too, since J4 is derived from J1 by valid arguments having only one other, tautological premise (J3). And so the protagonist would in this case, by proving J6, have undermined her own grounds for accepting J1. The moral of this story is that the protagonist can know she is consistent, and J1 can be true, only if the protagonist won't take the risk of lapsing into inconsistency involved in proving J6. In short, *J1 entails that the protagonist will not prove J6*, from which it follows immediately that J1, J3, J5 ⊢ J6 is a perverse argument. (Its premises entail both J6, and that the protagonist won't prove J6.) Thus J8 is false and the mathematical argument is unsound.

If the above analysis is right then the mathematical argument is valuable, not because it shows that the human mind's problem-solving powers exceed those of a machine (it doesn't), but because it provides a wonderful, non-contrived example of a case where one must reason reticently, rather than classically, to avoid succumbing to a perversity fallacy.

6. Conclusion

In this paper I have shown that CL exposes us to a kind of fallacy – the 'perversity fallacy' – wherein one accepts the conclusion of a classically valid argument even though its premises entail that one will not accept it, with the result that one falsifies the premises and undermines one's grounds for accepting the conclusion in the very act of accepting it. I have argued that we should instead reason in accordance with RL – a logic that includes a 'perversity check'. I have briefly sketched several versions of RL, and demonstrated that the notions of perversity and reticence have an important bearing on major issues in the philosophy of artificial intelligence.

Issues raised by this paper that for reasons of space I must save for future work include: (i) applying RL to analyzing Moorean sentences and what Sorrensen (1988) calls 'blindspots'; (ii) using RL to critique the doctrine that knowledge and/or justified belief is deductively closed; (iii) investigating 'higher-order perversities' (wherein the premises of an argument entail, not only that one won't accept the conclusion, but also that one won't detect the perversity); (iv) contrasting RL with other non-classical logics; and (v) further investigating the properties of weak and strong RL.

Acknowledgements

Many thanks to Thomas Forster, Matthew Grice, Doukas Kapantais and Michael-John Turp for comments and suggestions.

References

Barr, M. (1990). Review: The Emperor's New Mind. By Roger Penrose. *The American Mathematical Monthly,* 97(10), 938-942.

Benacerraf, P. (1968). God, the devil and Gödel. *The Monist*, 51, 9-32.

Boolos, G. (1990). On seeing the truth of the Gödel sentence. *Behavioral and Brain Sciences,* 13(4), 655-6.

Bowie, G. L. (1982). Lucas' number is finally up. *Journal of Philosophical Logic,* 41(3), 279-285.

Chalmers, D. (1995). Minds, machines, and mathematics. A review of Shadows of the Mind by Roger Penrose. *Psyche,* 2, 11-20.

Gaifman, H. (2000). What Gödel's incompleteness result does and does not show. *The Journal of Philosophy*, 97(8), 462-470.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monash. Math. Phys*., 38, 173-198.

Gödel, K. (1951). Some basic theorems on the foundation of mathematics and their implications. In Gödel (1995) (pp. 304-323).

Gödel, K. (1995). *Collected works, vol. III: unpublished essays and lectures.* (S. Feferman, et al., eds.) Oxford: Oxford University Press.

Grush, R. & Churchland, P. (1995). Gaps in Penrose's toilings. *Journal of Consciousness Studies,* 2(1), 10-29.

Lindström, P. (2001). Penrose's New Argument. *J. Philosophical Logic,* 30, 241-250.

Lindström, P. (2006). Remarks on Penrose's New Argument. *J. Philosophical Logic,* 35, 231-7.

Lucas, J. R. (1961). Minds, machines and Gödel. *Philosophy*, 36, 112-127.

Lucas, J. R. (1996). Minds, machine and Gödel: a retrospect. In P.J.R. Millican and A. Clark, eds., *Machines and Thought: The Legacy of Alan Turing*, Oxford: Oxford University Press. 103-124.

McCullough, D. (1995). Can humans escape Gödel? A review of Shadows of the Mind by Roger Penrose. *Psyche,* 2(23), 57-65.

Nagel, E. & Newman, J. (1957). *Gödel's proof.* New York: New York University Press.

Penrose, R. (1989). *The emperor's new mind.* Oxford: Oxford University Press.

Penrose, R. (1994). *Shadows of the mind.* Oxford: Oxford University Press.

Penrose, R. (1996). Beyond the doubting of a shadow. *Psyche,* 2(23), 89–129.

Putnam, H. (1960). Minds and machines. In Sidney Hook, ed., *Dimensions of mind: a symposium*. New York University Press. Reprinted in Anderson, A. R., ed., 1964. *Minds and machines*. Prentice-Hall, 77.

Raatikainen, P. (2014). Gödel's incompleteness theorems. *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition). Edward N. Zalta (ed.). http://plato.stanford.edu/archives/spr2014/entries/goedel-incompleteness/.

Robinson, W. (1992). Penrose and mathematical ability. *Analysis,* 52(2), 80-87.

Rucker, R. (1982). *Infinity and the mind: the science and philosophy of the infinite.* Princeton, N.J.: Princeton University Press.

Shapiro, S. (2003). Mechanism, truth and Penrose's New Argument. *J. Philosophical Logic,* 32, 19-42.

Sorensen, R. (1988) *Blindspots.* Oxford: Clarendon Press.

Turing, A.M. (1947). Lecture to the London Mathematical Society on 20 February 1947. Reprinted in D.C. Ince (1992), ed., *Collected works of A.M. Turing: mechanical intelligence,* Amsterdam: North Holland.

Turing, A.M. (1948). Intelligent machinery. Reprinted in D.C. Ince (1992), ed., *Collected works of A.M. Turing: mechanical intelligence,* Amsterdam: North Holland.

Turing, A.M. (1950). Computing machinery and intelligence. Reprinted in D.C. Ince (1992), ed., *Collected works of A.M. Turing: mechanical intelligence,* Amsterdam: North Holland.