Gödel, Penrose and Paraconsistency: What Goes? What Stays?

Kartik Tiwari*

May 14, 2023

1 Introduction

Though Gödel's Incompleteness Theorems (GIT) [1] are ubiquitous in metalogic philosophy discussions, they have also entered discourse on philosophy of mind primarily through two popular works. First, through Douglas Hofstader's 'Strange Loop Constructions' (introduced in 'Gödel, Escher, Bach' [2]) which he employs to explain an emergent self in the form of tangled hierarchies [3]. Then, through Roger Penrose's argument[1] for non-computability of human intelligence in 'Emperor's New Mind' [4] (which later got refined in 'Shadows of the Mind' [5]). While Hofstader's use of Gödel's results were mostly metaphorical in nature, in the Lucas-Penrose Argument they are placed right at the heart of the action. There have been many objections to the Lucas-Penrose argument which have received mostly dis-satisfactory responses and have led to a diminished interest in the subject amongst mainstream philosophers of mind. In stark contrast is the field of para-consistent formal systems which has been gaining increasing traction over the past decades. Naturally, one is led to an inquiry about the

*Department of Physics and Department of Philosophy, Ashoka University

[1]Note that a version of this argument was presented before Penrose by John Lucas. Even earlier versions could be found in the works of E. Nagel and J. R. Newman. Arguably, however, none were as popular as Penrose's. It is widely accepted that publication of 'Emperor's New Mind' revived an interest in Gödelian arguments on human intelligence.

status of Lucas-Penrose Argument and its objections in the light of paraconsistency. In this paper, we begin by briefly introducing the Gödel's (First) Incompleteness Theorem, Lucas-Penrose Argument and Paraconsistent Formal Systems. Then, we summarize - what is thought to be - an authoritative dismantling of the Lucas-Penrose argument by David Chalmers. Following this, we systematically review the status of some possible objections that arise in the context of paraconsistency. We conclude with some speculative remarks about paraconsistency in the broader discussions on human intelligence.

2 Assembling Pieces

In this section, we briefly introduce the technical pieces required before move on to the discussions on paraconsistency. Since the primary objective of this manuscript is to review the status of these results in paraconsistent frameworks, the introductions here are merely pedagogical.

2.1 Gödel's Incompleteness Theorem

Kurt Gödel's ingenious trick was finding a clever technique to get a formal system (like Peano Arithmetic) talk about itself. Before Gödel, Russell and Whitehead had hoped in their Principia Mathematica to axiomatize mathematics in a way that rids itself of paradoxes that arise due to self-reference (like the liar's paradox, Russell's paradox, etc.). What Gödel's proved was that any formal system that is as expressive as Peano Arithmetic (or more) is either inconsistent or incomplete. Heuristically, the argument relied on coding a Gödel sentence for a given formal system F that resembled the form

$$G(F) \equiv \text{This sentence has no proof in system F}$$

If the truth of Gödel sentence can be derived through a proof then there exists a proof of $G(F)$. This means we end up a contradiction with both $G(F)$ and $\sim G(F)$. On

the other hand, if there is no proof of $G(F)$ then $G(F)$ is true but without a proof. Therefore, we either get a system that is complete but inconsistent. Or, we get a system which is consistent but with at least one true statement that cannot be proved. There is of course a whole host of techincal nuance which makes such statements possible which we have skipped but these can be found easily in [1, 6, 7, 8, 4].

2.2 Lucas-Penrose Argument

But what has any of it to do with intelligence? This is where the Lucas-Penrose argument comes in. There have been different versions of the argument but, here, I (adapt and) summarize what is considered one of the clearest and most charitable expositions of Penrose's Second argument in 'Shadows of the Mind'[2] [5] by Chalmers [9]. At first reading, the argument seems slightly roundabout but we shall come back and clarify why it is structured the way it is.

1. **Assume** my reasoning powers follow some formal system F
2. Since I know I am F and I am sound, I know F is sound.
3. Consider another formal system F' that is the union of F with the assumption made in (1).
4. I know F' is sound because supplementing a sound system with a true statement yields a sound system still.
5. I know the Gödel sentence of F' (i.e. G(F')) is true because I know Gödel's Incompleteness Theorem is true and that F' is consistent.
6. Since F' is consistent, no amount of algorithmic reasoning within F' would lead to a proof of G(F').

[2]It could also be worth mentioning that Penrose constructs this argument regarding the inability of a formal system to capture human reasoning and using a theorem by Craig claims that human intelligence, and subsequently consciousness, is non-computable. However, currently known physics is indeed computable. He uses these two assertions to indicate that radically new physics is required to solve the mystery of consciousness.

7. I can 'see' the truth of G(F') but F' can never 'see' the truth of G(F').

8. But I am effectively F'. So, I should both know and not know that G(F') is true. This is a contradiction and we must discharge the assumption made in (1).

9. Therefore, F must not have captured my reasoning powers completely.

10. Since F was an arbitrary formal system, no formal system could have captured my reasoning powers completely.

Intuitively, one would think a more parsimonious way of arriving at this result would be to simply claim that if my reasoning is captured by a sound formal system F which cannot know the truth of its Gödel sentence G(F) then my knowledge of G(F) being true contradicts the initial assumption. However, this formulation admits objections about F being 'knowably sound'. These objections are discussed in [9] and are not very relevant to our current discussion. At this stage, we just note that the aforementioned formulation circumvents a few usual objections to the naive presentation of the Lucas-Penrose argument.

2.3 Chalmers' Soundness Objection

Chalmers' claims [9] that the contradiction in the above argument is not generated by the assumption that I am (algorithmically) isomorphic to the formal system F. Instead, the contradiction arises primarily in (2) where we claim unassailably that I am sound i.e. F is sound. Chalmers presents an adaptation of a Lob's Theorem in his article to support this objection. Here, we state directly the relevant result but a general proof can be found in [8]. Lob's Theorem states that any formal system that can prove its own soundness becomes inconsistent[3]. Therefore, even before the argument proceeds, in some ways we have already created an inconsistency by asserting that we know we are sound. Since Lob's Theorem is a powerful meta-mathematical results, this case was widely considered shut.

[3]This is a strange result but an intuitive exposition is provided by by Eliezer Yudkowsky in this cartoon (clickable link) on LessWrong

2.4 Paraconsistent Formal Systems

Paraconsistent formal systems [10, 11] were developed to discard the Principle of Explosion. In classical logics, the fact that we can perform reductio ad absurdum proofs starting with assuming the negation of a theorem and proving a contradiction to discharge the negation rests on the fact that contradictions render a classical formal system trivial (and we wish to avoid triviality at all costs). In other words, because of the Principle of Explosion, from a contradiction in classical logic any other statement can be proved. This leaves the logic useless for all practical purposes should a contradiction arise in classical logic. The notion of paraconsistency was introduced to coherently reason about inconsistencies without exploding the system into triviality. Paraconsistent systems cordon off inconsistencies in a way that their logical entailments do not wildly travel to domains where we do not want them to. This opens up possibilities to construct formal systems that aspire to more accurately model linguistics, scientific reasoning, causal modus ponens, adaptive reasoning, paradoxes, etc. The key take-away in this subsection is that there are non-trivial paraconsistent logical systems in which, over relevant domains, both a statement and its negation can be true.

3 Revisiting Lucas-Penrose under Paraconsistency

One attempt to perform a similar analysis as that of the current was done by Megill [12]. However, I find the argumentation presented in [12] loose, incomplete and unsatisfactory. Megill uses paraconsistency to simply dissolve Lucas-Penrose's arugment into a weaker form that does not concludes 'human intelligence is non-computable'. Instead the weaker form concludes the disjunction 'either human intelligence is non-computable or it follows a paraconsistent formal system'. In this section, I systematically study which objections to Lucas-Penrose argument still carry weight after inclusion of paraconsistency and which ones fail to get off the ground. Having introduced the relevant technical pieces, let us review the status of the some objections to Lucas-Penrose Ar-

gument and understand which ones are admissible in paraconsistent frameworks and which ones are not.

3.1 Objection 1 - Revisiting Soundness Attack

Chalmer's objection from Lob's theorem on premise (2) was about the fact that if F can assert its own soundness, then F has to be inconsistent. If a paraconsistent formal system has to be maintained as non-trivial, then the proof of Lob's Theorem fails (as argued by Priest in [8]). Therefore, now it does not necessarily follow that a theory which asserts its own soundness is inconsistent. Note, though, we are working with a paraconsistent system by assumption. Recall that paraconsistency differs from inconsistency (at least) in the fact that paraconsistency does not render the formal system trivial. In classical logic, consistency is a necessary condition for a system to be sound. However, it is possible to construct a non-trivial paraconsistent system that can accommodate contradiction without sacrificing soundness. Thus, it is (logically) unproblematic to assert premise (2) as long as F is a paraconsistent formal system. Chalmer's objection using an adaptation of Lob's theorem fails in paraconsistent regimes.

3.2 Objection 2 - Proof by Contradiction

However, one is now led to the concern that the structure of Lucas-Penrose argument follows a classic *reductio ad absurdum*. Therefore, if paraconsistency is considered, then the Lucas-Penrose argument would not even get off the ground. This is also not true because it is indeed possible to construct paraconsistent formal systems which allow for *reductio ad absurdum* proofs from certain classes of contradictions (relevance logic is an example of a paraconsistent system that admits proofs by contradiction). This assertion follows from the fact even though paraconsistent systems accept contradictions, they need not accommodate *all* contradictions. Rather, paraconsistent systems are designed to be able to study logical consequences of contradictions in a

controlled manner.

3.3 Objection 3 - Special Status of $G(F)$

Even though paraconsistency lets Lucas-Penrose argument avoid the previous two objections, it is a feature of paraconsistency itself that provides an inescapable objection which dissolves the argument too. First recall that G*ö*del's incompleteness theorem does not state that sufficiently expressive formal are incomplete - rather, it states that sufficiently expressive formal systems are either incomplete or inconsistent. Since we usually tend to assume consistency, it is the completeness that we compromise on. In paraconsistent systems, however, we accommodate (partial) inconsistency but that lets us escape incompleteness. To understand this, observe what happens to the G*ö*del's sentence in a para-consistent system. If G(F) does not have a proof, then $G(F)$ is true. Alternatively, if $G(F)$ has a proof,then we end up with both $G(F)$ and $\sim G(F)$ as described in §2.1. In paraconsistent regimes, this is not a problem because we can construct our system such that it accommodates this contradiction without crumbling the rest of the logic into triviality (see §2.4 [**?**]. In some ways, then, the G*ö*del sentence in paraconsistent systems loses its 'special status' because even after analysing the $G(F)$ the system remains complete and paraconsistent. Another way to think about this is to recognize that at the heart of G*ö*del's proof is a paradox that we try to avoid but paraconsistent logics can comfortably accommodate. This is a concern for Lucas-Penrose argument because it proceeds towards the conclusion about non-computability of human intelligence on the basis of the assertion that the formal system F' would never know the truth of $\sim G(F)$ because in the classical case $G(F')$ would have lacked a proof. Since the paraconsistent system is complete, if F is allowed to be paraconsistent then F' could indeed prove the truth of $G(F')$. This resolves the inconsistency between human intelligence both knowing and not knowing the truth of $G(F')$.

4 Mechanisms in Light of Paraconsistency

In the last section, we established that - under paraconsistent considerations - the authoritative argument by Chalmers against Lucas-Penrose fails and that the *reductio ad absurdum* structure of Lucas-Penrose could justifiably be maintained but the argument still fails to conclude the uncomputability of human intelligence because the Gödel sentence loses its special status. It is important to note, however, that the objections presented in this paper undercutting defeaters and not rebutting defeaters. Therefore, we can only conclude that even if human intelligence is non-computable, Lucas-Penrose argument fails to explain why. Mechanism, that is the philosophical doctrine which considers all living beings to be complicated machines, is neither supported nor attacked on the basis of Lucas-Penrose argument alone. Without allowance of paraconsistency, Lob's theorem fails the Lucas-Penrose argument and with allowance of paraconsistency, provability of Gödel sentence fails it.

Though it seems unlikely that a definitive argument against mechanism can be derived from purely formal considerations, the (relatively) new domain of computational complexity theory holds some promise [13]. Comparing estimates of the upper bounds on the computational competence of the brain and the lower bounds on the time complexity of feats of human intelligence certainly seems like a low-hanging fruit that could provide sweet results and serve as a seed for various exciting future investigations.

References

[1] Kurt Gödel. On formally undecidable propositions of principia mathematica and related systems.

[2] Douglas R. Hofstadter. *Gödel, Escher, Bach: an eternal golden braid*. Basic Books, 20th anniversary ed edition.

[3] Douglas R. Hofstadter. *I am a strange loop*. Basic Books.

[4] Roger Penrose. *The emperor's new mind: concerning computers, minds and the laws of physics*. Oxford landmark science. Oxford University Press, revised impression as oxford landmark science edition. OCLC: ocn948795136.

[5] Roger Penrose. *Shadows of the mind: a search for the missing science of consciousness*. Oxford Univ. Press, 1. paperback ed edition.

[6] Panu Raatikainen. Gödel's incompleteness theorems. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2022 edition.

[7] Dag Westerståhl. *Foundations of logic: completeness, incompleteness, computability*. CSLI Publications & Tsinghua University.

[8] Graham Priest. Löb's theorem and curry's paradox. 6(3). Section: Logic.

[9] David Chalmers. Minds, machines, and mathematics.

[10] Graham Priest, Koji Tanaka, and Zach Weber. Paraconsistent logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2022 edition.

[11] Stewart Shapiro. Incompleteness and inconsistency. 111(444):817–832.

[12] Jason L. Megill. Are we paraconsistent? on the luca-penrose argument and the computational theory of mind. 27(1):23–30.

[13] Scott Aaronson. Why philosophers should care about computational complexity.