OPEN ACCESS
Journal of Artificial Intelligence and Consciousness
Vol. 9, No. 2 (2022) 241−266

DOI: 10.1142/S2705078522500011

World Scientific
www.worldscientific.com

Douglas Hofstadter's Gödelian Philosophy of Mind

Theodor Nenu

Department of Philosophy, University of Bristol
Cotham House, Bristol, BS6 6JL, UK
theodor.nenu@bristol.ac.uk

Received 13 December 2021
Revised 14 February 2022
Accepted 16 February 2022
Published 9 April 2022

Hofstadter [1979, 2007] offered a novel Gödelian proposal which purported to reconcile the apparently contradictory theses that (1) we can talk, in a non-trivial way, of mental causation being a real phenomenon and that (2) mental activity is ultimately grounded in low-level rule-governed neural processes. In this paper, we critically investigate Hofstadter's analogical appeals to Gödel's [1931] First Incompleteness Theorem, whose "diagonal" proof supposedly contains the key ideas required for understanding both consciousness and mental causation. We maintain that bringing sophisticated results from Mathematical Logic into play cannot furnish insights which would otherwise be unavailable. Lastly, we conclude that there are simply too many weighty details left unfilled in Hofstadter's proposal. These really need to be fleshed out before we can even hope to say that our understanding of classical mind-body problems has been advanced through metamathematical parallels with Gödel's work.

Keywords: Gödel's Incompleteness Theorems; Self-consciousness; Mental Causation; Strange Loops.

1. *Gödel's Theorem and The Brain*: Introduction

Philosophical literature brims with proposals which make use of Gödel's Incompleteness Theorems in order to reject computational accounts of human minds (e.g. Nagel and Newman [1958], Lucas [1961], Penrose [1989]). One may rightfully get the impression that the very same theorems cannot possibly be used for philosophically opposed purposes, e.g. to philosophically illuminate computational perspectives on the mental. That being said, such "positive" Gödelian theses exist, but for some mysterious reason, they have generated virtually no engagement in the academic discourse. This paper aims to make some steps toward filling this literature gap by focusing on a specific positive proposal: that of Hofstadter [1979, 2007].

In his tomes, Hofstadter offers an elaborate thesis about how inanimate components (such as individual neurons) can collaborate to yield a conscious mind with causal powers. Our aim in this paper is to look at just a handful of aspects of his larger project, such as his explanation of why it is perfectly meaningful to talk of human minds as being causal *loci*, whilst still acknowledging the deterministic nature of the low-level neural computation which takes place in human brains. Hofstadter maintains that his insights are best observed through parallels with the Gödelian Incompleteness Phenomenon in Metamathematics.

The task we set for ourselves in these pages is to take a closer look at these parallels and to see what lessons can we actually derive from them. Originally, Hofstadter's magnum opus, "Gödel, Escher, Bach" (GEB) was planned to be a much shorter piece, suggestively entitled "Gödel's Theorem and The Brain". To see what The Incompleteness Theorems supposedly have to do with brains (and with questions of causality), we will answer within this paper the following questions:

(1) What does Hofstadter mean by a Strange Loop (or a Tangled Hierarchy)?
(2) Why does Hofstadter invite us to view the brain from various hierarchical vantage points (e.g. the microphysical level, the neural level, the symbolic level, the level of the mind)?
(3) How does causality interplay between these hierarchical levels? Is the hierarchy a tangled one?
(4) What quintessential Strange Loop does Hofstadter think Gödel's work contains?
(5) How, why and in what way could that particular Strange Loop illuminate the way in which we understand minds?

2. What is a Strange Loop?

The phrase "Strange Loop" was coined by Hofstadter [1979]. As it is the case with most categories, it best understood through illustrative examples, rather precise definitions in terms of necessary and sufficient properties. There are two main kinds of strange loops and we will offer a prototypical example of both. Before we do so, we quote a tentative definition provided by Hofstadter [2007]:

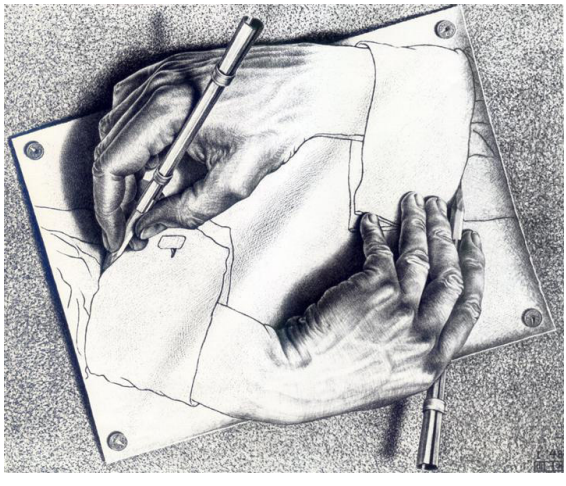
> What I mean by "strange loop" is (. . .) an abstract loop in which, in the series of stages that constitute the cycling-around, there is a shift from one level of abstraction (or structure) to another, which feels like an upwards movement in a hierarchy, and yet somehow the successive "upward" shifts turn out to give rise to a closed cycle. (p. 102)

This way of putting it may sound complicated and abstract. Here is our first famous example of a strange loop which should shed some light on the phenomenon. Consider the following sentence, also known as The Liar Sentence:

This sentence is not true.

Is that sentence true? If it is, then it asserts a truth, namely that it is not true, which contradicts the initial assumption. Hence, it cannot be true. But if it is not true, then it conforms to what it says, so it is true after all! We are thus caught in a strange loop where in the reasoning stages, we seem to make a bit of upward progress in the chain of deductions, only to get back to the initial point. In short, the strange loop phenomenon manifested itself.[1] In his proof of The First Incompleteness Theorem, Kurt Gödel used a "liar-like" trick: not with the semantic notion of *truth*, but with the syntactic notion of *formal provability*. For Hofstadter, that proof contains a prime example of a *genuine* strange loop. We will say more on this in Sec. 5.

For the mind-brain discussions that will follow, we need one more example of a strange loop: an illusory strange loop. I will pick one that is cherished by Hofstadter, authored by Escher, the Dutch artist whose work usually abounds with strange loops. Escher's [1948] "Drawing Hands" is a canonical example. When one ponders what is depicted by it, a paradoxical hierarchical structure engendered by the drawer-drawn distinction emerges.



The strange loop goes as follows: the left-hand appears to be drawing the right hand, giving the impression that it sits *higher* in the hierarchical structure generated by the picture (because it is the drawer). But one will quickly notice that the existence of the left hand is owed to being drawn by the right hand, so it must sit *lower* after all. An endless cycle ensues.

However, as the reader can surely anticipate, there is an obvious "fake" aspect of this strange loop, namely that it is not genuinely paradoxical — unlike, say, the

[1] One may be tempted to say that the Liar sentence is neither true nor false: so it is (i) not true and (ii) not false. In particular, given (i), it is not true. But that is what it says! We're back in the loop...

strange loop of The Liar Sentence. The paradox is just merely apparent: it only takes place at the level of the drawing. The drawer−drawn relationship between the hands is indeed artfully depicted on paper, but that particular hierarchical structure is not *actually* instantiated in reality (regardless of the fact that a drawing attempt of this paradoxical state of affairs exists on Escher's desk). In a more poetic way, we can say that the universe will not explode because Escher drew his famous "Drawing Hands", for he cannot generate with the aid of a pencil an actual metaphysical crack in the fabric of reality. The important takeaway is this: there is an *invisible level* (or inviolate level) behind the *Drawing Hands* loop (namely the external one where Escher is situated). The inviolate level does not exhibit any loops whatsoever, only the level of the drawing does.

Thus, strange loops come in two main flavors: genuine strange loops and apparent strange loops. We can also call them *Escher-like Strange Loops* and *Gödel-like Strange Loops*. Hofstadter will ultimately want to argue that the hierarchy of brain levels is an Escher-like tangled hierarchy or strange loop, but that we can nonetheless import some Gödelian lessons in our understanding of minds.

As we just said, in the Drawing Hands strange loop, there is an inviolate level where Escher resides: there the paradox evaporates. Likewise, the level of neurology is an inviolate level in the Escher-like strange loop of consciousness: only the level of the mind, which is situated higher in the hierarchy, exhibits entanglement. Hofstadter [1979] points out that we may mistakenly be under the illusion that there is no inviolate level, but this would be erroneous:

> The illusion is created, because of the Tangled Hierarchy of symbols, that *there is no inviolate level*. One thinks there is no such level because that level is shielded from our view. (...) This is an interesting case where a software tangle, that of the symbols, is supported by a hardware tangle, that of the neurons. But only the symbol tangle is a Tangled Hierarchy. The neural tangle is just a "simple" tangle. (p. 691)

In the following section, we will offer more details on what symbols and self-symbols are, together with outlining the relationships between them and our biological neural network. In Sec. 4 we will present "the elements of strangeness" behind a particular mentalistic strange loop, that of mental causation. By the start of Sec. 5, we will have everything in place to properly explore what Gödel's work has to do with all of this.

3. Particles, Neurons, Symbols and Minds

This section will fill in some important details about the main brain levels which are usually targeted in Hofstadter's explanations. In particular, we will present what Hofstadter has in mind when he talks of the symbolic level of the brain, but we will also touch on the level of the mind, which is a higher hierarchical level where the self-symbol (which corresponds to consciousness) is postulated to reside.

The foregoing talk about symbols and neurons naturally sets up the stage for a larger theme of tackling the relationship between complex entities and their components. A big part of Hofstadter's philosophy is to account for higher order ontology in a world where, at fundamental physical levels, we cannot find any of the entities that we are so familiar with (e.g. cats, coffee mugs, etc.), let alone symbols. The language of microphysics prohibits reference to virtually all things that are most familiar to us. One of Hofstadter's aims is to explain why our lexical choices target macro-level categories instead of microphysical entities (and also to show why *mentalese* is wholly unlike microphysical languages).[2]

Hofstadter, like Dennett [1991] construes macro-objects (such as biological organisms, including us) as *complex patterns*. By restricting ourselves to the level of microphysics, we miss out on a lot of real structural facts about the systems that we want to analyze. Wallace [2012] embraces Dennett's insights and makes the following comment which captures the Hofstadterian spirit:

> (T)here are structural facts about many microphysical systems which, although perfectly real and objective (try telling a deer that a nearby tiger is not objectively real), *simply cannot be seen if we persist in describing those systems in purely microphysical language*. Zoology is of course grounded in cell biology, and cell biology in molecular physics, but the entities of zoology cannot be discarded in favour of the austere ontology of molecular physics alone. (pp. 48−49, my emphasis)

Speaking of Biology, we remark that Hofstadter's ontological inklings stem from a handful of evolutionary insights which support the idea that some patterns are, in a sense, more real than others. Darwinian processes are taken to have engineered important aspects of our brains which can *only* be seen by zooming out from the level of neurology. Brains of evolved creatures will turn out to be store-houses for many highly interesting patterns called *symbols* which mirror reality — or, more precisely, the reality that we find ourselves in, viz. macro-level reality.

Symbols are brain structures which correspond to concepts or categories. Hofstadter motivates their existence by remarking that natural selection is particularly unforgiving of creatures that underperform at classifying various bits of their environment, especially those which directly impact upon survival prospects (e.g. food, predators or mates).

Thus, minds of evolutionarily successful species are expected to be bestowed with categorization abilities tuned for picking up on various patterns. A curiosity ensues: a constant element of any creature's environment — that is an entity that is always there — *is the creature itself*, so at least a basic self-categorization is expected to occur [2007, p. 74]. This kickstarts the formation of the self-symbol, whose details we will be unraveled toward the end of this section and in Sec. 4.

[2]Hofstadter does not use the word *mentalese* or explicitly talk about Fodorian *languages of thought*, but there is little doubt that he subscribes to some species of the idea, given his account of thinking that we will present in a moment.

For now we should come back to ordinary symbols. The word "symbol", as we said in the previous paragraph, is Hofstadter's lexical choice for the hardware realizer of concepts. What is the physical architecture of a symbol? How do they implement concepts? Hofstadter is agnostic about their "shape": he simply takes symbols to consist of bundles of neurons which fire together in specific patterns. This synchronized firing (which *activates* the symbol) will take place when the being in question thinks about or directly perceives an instance of the concept that the symbol is implementing. Hofstadter is cautiously vague about the more specific details of symbols, since those details are largely irrelevant for his purely metaphysical views about mental causation and might engender discussions that could have been easily avoided.

We shall now isolate some aspects of symbols which we will make use of in Sec. 5 where we will argue against the possibility of talking about a "psychological variant" of the Incompleteness Theorems to which symbolic-networks (or "mental systems") can be subjected.[3] An *activation* of a *dormant* symbol corresponds to synchronized firings of the individual neurons which mereologically sum up to that symbol. As it can be guessed, the intensity of the activation inside the neural complex depends on the firings of the individual neurons which constitute that symbol, and hence symbol activation is something which comes in degrees.[4] Another thing worth pointing out, but which perhaps can be guessed by our focus on *patterns* of neural firings, is that symbols are not clearly delineated entities and, also, lots of symbolic overlapping takes place.

We know that most living things think about things, but what is thinking? First of all, we remark that, for Hofstadter, thinking is an activity which happens not at the level of symbols, but at the level of the mind. On his picture, a good metaphor which captures the essence of thinking is the following: thoughts are basically *trips* through symbols. On the journey, various dormant symbols are woken up to different intensity levels. Here is a way of imagining the mental map and the level of the mind:

> If it were possible to schematize this whole image, there would be a gigantic forest of symbols linked to each other by tangly lines like vines in a tropical jungle — this would be the top level, the Tangled Hierarchy were thoughts really flow back and forth. This is the elusive level of *mind*. (1979, p. 691)

Now we come back to the mismatch between the language of microphysics and English (or any other natural language), which seem to reflect rather distinct ontological baggages. Why is that? When we speak, we couch our thoughts in linguistic

[3]Important: Hofstadter *does not claim* in *GEB* that there is a variant of the Incompleteness Theorems for mental systems, but many misunderstood him as implying that.

[4]This is not entirely surprising, given well-known work in the Psychology of Concepts on *typicality effects*. For a comprehensive overview of typicality, see Murphy [2002, Chap. 2]. We however remark that there is no evidence which vindicates typicality at the symbolic level of the brain. That being said, to my knowledge at least, there is also no neurological evidence which refutes it.

form. Since thoughts are trips through symbols, then there is no surprise that our choices of lexical tags will reflect macro-level reality, for this is the reality which gets mirrored in our symbolic network.[5] We are macroscopic structures and the waters in which we swim are those revealed by the so-called Manifest Image [Sellars, 1963]. Here we find things such as trees, balls, dogs or colors — but we do not find entities such as quarks, electrons or ribosomes. Education can make us know about these, of course, but we are forever doomed to *intuitively process reality* at our level (and to consider this level as the realest there is).

To recap, we arrive at the symbolic level of the brain by zooming out from the level of neurology until we allow ourselves to notice some real, objective structural facts about biological neural networks.[6] One can zoom out even further: this time from the level of symbols to the level of mind, the latter being the only level where the most special and complex symbol of the brain, the self-symbol, can be noticed and appreciated. In his earlier work, Hofstadter calls the self-symbol a *subsystem* (we will also sometimes talk of *the pattern of selfhood*, but one can switch back and forth between these expressions).

Subsystems are still symbols, for they are abstract high-level neural patterns, but they host plenty of interacting "ordinary" symbols. To confirm, there are many subsystems in the brain, but the one which correspond to the self is the main player and the most complex of them all.[7] Hofstadter [1979] deems consciousness to be a mirage which is a property of this privileged complex symbol:

> A very important side effect of the *self*-subsystem is that it can play the role of "soul", in the following sense: in communicating constantly with the rest of the subsystems and the symbols in the brain, it keeps track of what symbols are active, and in what way. This means that it has to have symbols for mental activity—in other words, symbols for symbols, and symbols for the actions of symbols. (...) (D)espite its earthly origin, this way of describing awareness—as the monitoring of brain activity by a subsystem of the brain itself—seems to resemble the nearly indescribable sensation which we all know an call "consciousness". (pp. 387–388)

The existence of the self-symbol is again not explicitly philosophically argued for by Hofstadter. In Sec. 5, however, we will encounter some *quasi*-Gödelian parallels that he makes between how brains acquire a self and how formal systems acquire a

[5]The level targeted by our symbolic structure is one that is optimal for allowing the creature whose brain hosts these symbols to maximize comprehensibility and predictive power. For example, tigers are more real for deers than anything else because the symbol which mirrors the tiger-pattern is crucial for survival prospects.

[6]This is largely the same thing done in physics when we move from statistical mechanics to thermodynamics.

[7]It is important to note these symbols that are hosted by the self-subsystem at any time are not just the ones which are directly triggered through vision or other sensory input. Among the hosted symbols are also symbols drawn from the storehouse of Episodic Memory, i.e. specific memories of our life which contribute to both "I"-ness and personal identity.

"self". But it is extremely important to note that this will not pretend to be a real ontological defence of the existence of a self-symbol — the brain subsystem that corresponds to selfhood is basically assumed to exist. The task Hofstadter sets for himself in his work is to shed some Gödelian light on various mysterious properties that philosophers tend to be interested in. We will expand on this when the time comes, but we end by saying that whilst Gödel left plenty of instructions as to how to construct a Gödel sentence for *Principia Mathematica*, Hofstadter left no instructions for constructing a self-symbol out of a complex neural network.[8]

4. The Strange Loop of Mental Causation

Sperry's [1965] paper moulded Hofstadter's intuitions on the issue of mental causation. Sperry believed that there is a lot of high-order emergence which takes place "in the head" of a person and that there are plenty of neural firing orders issued from "higher-order commands". Hofstadter felt that Sperry insightfully intuited that the question of "who pushes whom around in the population of causal forces which occupy the cranium" is not by any means straightforward. Hofstadter's analysis of mental causation through strange loops is his own attempt to settle his own variant of Sperry's question.

Because Hofstadter equates thinking with consciousness[9] and because the strange loop of mental causation will turn out to be an Escher-like strange loop owed to the hardware limitations of our symbolic structure, one can see Hofstadter's proposal as a purported answer to Chalmers' [2018] *meta-hard* problem of consciousness. The standard hard problem of consciousness asks why and how can neural computation or brain processes give rise to subjective experience, the meta-hard one concerns our puzzlement about these issues: why are we so phased by mind-body problems?

To present Hofstadter's explanation of why we are bound to feel ourselves as causal *loci*, we recall that human beings are macro-structures evolved by natural selection to perceive reality at their own level only. These instincts are so hardwired into us and culturally entrenched that no amount of education can override them. We may learn a lot about fundamental particles, DNA molecules and other lower-leveled entities, but they cannot have the same *reality* to us. Or, as Hofstadter puts it, too much (inevitable) brain-washing has taken place.

A *seed* of the self-symbol is there when we are born, but we cannot really talk of infants really possessing a (proper) self-symbol. Our brain lacks one at early stages, but with every moment of time that passes the self-symbol grows and grows until it eventually becomes the most overgrown symbol in the brain. We interact with the

[8]However, depending on the spirit in which one approaches the problem, these details might not be needed to appreciate the subtle points that he will make. For a neuropsychologist, these architectural details might matter more than they do for a philosopher interested in the purely conceptual issues involved.

[9]The fact that Hofstadter identifies between thinking, consciousness, having semantics, intentionality, having a soul and other notions that philosophers are interested in is unambiguous. For example, in his *Metamagical Themas* (Chap. 26) he explicitly says on the first page that he thinks all these phrases are synonymous.

world, we see other macro-patterns move when we interact with them and we gradually lock-in the illusion that *we* (viz. the macrostructure constituted by us) have causal powers which, at least apparently, allow us to make some other macro-structures in the environment react in various ways (e.g. this ball flew in the air after it was *kicked* by my leg).

Due to our representational universality,[10] our self-symbol grows with time because of the non-stop feedback loop maintained through constant interaction with the world. Quite importantly, when it comes to human beings, this feedback loop also has a *social dimension* whereby we seem to generate reactions (laughter, anger, flirts, etc.) in other *animate objects* (i.e. people).[11] Language also kicks in and our choice of words for macro-level objects make it the case that people can suddenly start referring to *us* through indexicals such as "you". This gradually shapes our "I"-ness. We maybe learn that we are smart, or ugly, or funny, or untalented, or terrible at philosophy, or bad at public speaking and so on. We internalize the lessons derived from the on-going feedback loops with the environment and the community and, through repeated trial and error, we are locked into a process which sharpens and grows the self-symbol with time:

> Those reactions bounce back to me and I perceive them in terms of my repertoire of symbols, and in this way I indirectly perceive myself through the eyes of others. I am building up my sense of who I am in others' eyes. My self-symbol is coalescing out of an initial void. (2007, p. 184)

This self-consciousness is an abstraction which is bound to have an undeniable reality for human beings, whose inner life is owed to "the dance of symbols" and to our quest of understanding the causal structure of the macro-world as finite beings that navigate through it. In this quest, self-hood is a reliable and indispensable emergent phenomenon that cannot be washed away, which unavoidably (out of *incorrigible ignorance* of the fundamental physical nature of reality) comes with the illusion that it is imbued with causal powers in the world of macroscopic patterns. In sum, the "I" is a useful linguistic and symbolic shorthand for this important macro-structure that is always there in our environment and which appears to have causality in the Manifest Image. Hofstadter [2007] sums up what we said so far as follows:

> You make decisions, take actions, affect the world, receive feedback, incorporate it into your self, then the updated "you" makes more decisions, and so forth, round and round. It is a loop, no doubt — but

[10] Hofstadter uses this phrase to single out the peculiar property of our symbols which allows for complex nestings which enable us to possess, in principle, almost any concept. He postulates that lower animals cannot do this: they could never acquire the concept of "online shop", for example.

[11] This sharpening of the self-symbol through social feedback loops is an idea which did not appear in Hofstadter [1979], but only in his more recent Hofstadter [2007]. The broad outline of this idea is spelled out at greater length by other contemporary cognitive scientists, e.g. Bogdan [2010]. The last section will talk more about the positive reception of Hofstadter's ideas.

> where's the paradoxical quality that I've been saying is a *sine qua non* for strange loopiness? (p. 193)

Why is mental causation a strange loop? Or, first of all, what are the supposed "ingredients of strangeness"? The main player in the strange loop of mental causation is that our cognitive architecture allows us to think. Thought, Hofstadter [1979, p. 337] says, "must depend on *representing reality in the hardware of the brain*". By now, we know how reality is mirrored in the hardware brain: through symbols, namely triggerable structures consisting of bundles of teamed up neurons which correspond to categories found at our level of reality.

Because thinking amounts to taking trips through symbols, it means that human thought can be as sophisticated as our symbolic network permits and the reality that it reflects. Since symbols only mirror the reality of the Manifest Image, a slight human handicap is revealed, for we automatically lack the required hardware to peer into lower-levels of the brain (and the world). This inability of ours to see what happens when we get to more fundamental brain levels — together with our ability to think — jointly facilitate the illusion that we are causal *loci*.

We are incapable of *truly* grasping that we are nothing more than epiphenomena yielded by lower-level goings-on. Even if we learn about these things in Philosophy or Neuroscience classes, the impulse to treat the self as the most real thing in the world and as a causal locus cannot possibly be eliminated.

Now it is easy to see why the hierarchy of brain levels exhibits a strange loop which pertains to causality. The Sperry question of who pushes whom around inside the cranium seems to have bottom-up answers when we move from fundamental levels such as that of microphysics to that of individual neurons. Similarly, we have bottom-up preservation of causality from the level of neurology to that of symbols. However, when we zoom out even further from the symbolic level to the level of the mind which hosts the self-symbol, it (mistakenly) seems to us as though causality flips, having a top-down nature from the self-symbol to lower-levels, yielding a loopy cycle. We recall that the strange loop of mental causation is an Escher-like strange loop and that this "downward causality" is ultimately illusory.

This is where Gödel's work is supposed to come in. Hofstadter believes that we would undersell mental causation if we left it like that, for he thinks that there are non-trivial and highly interesting senses in which mental causation can truly be said to be *real*. Human beings represent reality in the hardware of their brains and formal systems represent domains of *mathematical reality* in their symbolisms (Hofstadter [1979, p. 337]). His plan is to make us see downward causation with fresh eyes in the metamathematical setting and to import some of the lessons into the mentalistic setting. We now turn to see how this project can possibly work.

With the risk of repeating ourselves, we stress for clarity that Hofstadter is ultimately a reductionist [2007, Chap. 20] and that he perfectly understands that the sense in which mental causation is "real" is quite novel, abstract and, in a way, a fairytale. He would be the first to agree that physical causality is the primary causal

player — however, as he argued at length, the human-level view of the world cannot really be shaken off and this high-level perspective allows us to notice a "sort of" causal interplay between abstract patterns of physical entities. These interactions are full of statistical regularities which cannot be seen from the fundamental levels. Instead of being pessimistic about the perspective where abstract patterns can be appreciated, Hofstadter embraces this viewpoint and he attempts to squeeze out in human-language some interesting insights pertaining to the mind-body relationships which cannot be articulated in the language of microphysics. The strange loop of mental causation is Escher-like: this might mean that at its core it is not genuine — but this does not mean that it is devoid of insights.

5. Gödel's Theorem and the Brain

5.1. *What role is Metamathematics supposed to play?*

For Hofstadter, the work of Gödel contains the key behind understanding both consciousness and the reality of mental causation. We will see that it plays a *purely analogical role*, but this is not just any ordinary analogy. The relationship between formal systems and their Gödel-like sentences is supposed to be one of the very few concrete examples on which we have a firm grip and which can instill the prerequisite intuitions for understanding minds. In this section, we wonder whether bringing Mathematical Logic into play (in the way Hofstadter does it) can actually shed *extra* light on our current understanding of mind-body problems.

Self-consciousness has plausibly something to do with self-reference. To set the scene, Hofstadter [1979, 2007] points out to fruitless historical joint effort of Russell and Whitehead to keep self-reference at bay in their *Principia Mathematica*. Kurt Gödel's work effectively showed that if one purports to design a system that has at least a modicum of strength, then banning self-referentiality is an impossible task. As the reader can surely guess, similar self-referentiality is expected to occur in brains of sufficient complexity.

What does Hofstadter have in mind when he talks of Gödelian strange loops? The answer is that he is thinking about the so-called "diagonal construction" of the canonical fixed points of arithmetical predicates.[12] A fixed point of any unary arithmetical predicate $\phi(\mathsf{x})$ is an arithmetical sentence λ that is equivalent to $\phi(\overline{\ulcorner\lambda\urcorner})$. It can be shown that *every* predicate (pertaining to formal systems to which the

[12]The technique of Gödel-numbering allows us to assign numerical codes to syntactic objects pertaining to the formal system under investigation. For example, let's pick as our syntactic object an arithmetical sentence such as the following: $(\mathsf{0}+\mathsf{0})=(\mathsf{0}\times\mathsf{0})$. This sentence is at its core nothing more than a syntactic string of arithmetical symbols: $(,\mathsf{0},+,\mathsf{0},),=,(,\mathsf{0},\times,\mathsf{0},)$. Gödel showed how to encode any arithmetical string whatsoever via some natural number. The details of the coding scheme are not important, we simply say that the code of a sentence δ is usually written as $\ulcorner\delta\urcorner$. Thus, if δ is a syntactic string of characters, then $\ulcorner\delta\urcorner$ is a natural number among $\{0,1,2,3,\ldots\}$. We shall write $\overline{\mathsf{n}}$ as a shorthand for the object-linguistic arithmetical term consisting of n applications of the successor function symbol to arithmetical constant $\mathsf{0}$.

incompleteness theorems apply) has at least one fixed point. This important result of Mathematical Logic carries the name "The Diagonal Lemma".[13]

Let us pick a relevant formal system such as Peano Arithmetic. Gödel's merits were to ingeniously engineer a carefully designed predicate $\mathsf{Prov}_{\mathsf{PA}}(\mathsf{x})$ which applies precisely to the numbers which encode PA-theorems. Given the Diagonal Lemma, we can extract a fixed-point G_{PA} ("The Gödel Sentence of PA") of the negated predicate $\neg\mathsf{Prov}_{\mathsf{PA}}(\mathsf{x})$. Such Gödel-sentences can be mathematically shown to be formally undecidable, i.e. impossible to be formally proved or formally refuted starting with the PA axioms and toying from there with PA's deductive apparatus.

The fact that for relevant systems F their Gödel sentence G_{F} is formally undecidable is an undisputable mathematical fact, not a matter of philosophical interpretation. But there are potential philosophical interpretations of what is happening there in the diagonal proof. Hofstadter's colorful interpretation is that such diagonal constructions inject some *high-level meanings* inside sentences like G_{F}. This is done in a way which enable their mother system to talk about itself or, metaphorically speaking, gain a self. These high-level meanings will be an important player in this section.

Formal systems are frozen, rigid syntactic objects that are devoid of meaning. Brains are too meaningless syntactic entities, but this is only the case *if one views them at the neural or symbolic levels*, Hofstadter maintains. What imbues both mental systems and formal systems with meaning are the sophisticated isomorphisms that *inevitably* crop up when the complexity of the syntactic entity under consideration exceeds some relevant threshold: (1) for formal systems, we roughly need to be able to carry out elementary arithmetic inside them; (2) for brains, it suffices if they are complex enough to serve as media for hosting patterns that mirror the external world. Likewise, Gödelian self-reference is attained through customary diagonal methods; mentalistic self-reference, on the other hand, is achieved through the unavoidable entanglements that are obtained when the brain patterns which mirror the world *start mirroring themselves*.

Both the (supposedly) causally potent high-level meanings of Gödel-sentences and the causal potency of the self-symbol are intrinsically high-level phenomena, which means that they can only be understood at the appropriate high-level of analysis of the relevant syntactic object. This means that we cannot understand these phenomena at lower levels (or, even worse, we would not even be able to acknowledge them at all from these low vantage points).

Thus, we would miss out on the causes laying behind G_{PA}'s undecidability if we failed to properly reflect upon its high-level meaning. This meaning produces an upside-down flip whereby the quest for searching for bottom-up proofs (typically brought about by toying with axiom and rules) is automatically stopped and — in a

[13] Thus, by the "the diagonal construction of fixed points", we basically refer to the canonical proof underpinning The Diagonal Lemma. We cannot flesh it out here, but any good logic textbook contains it. Hofstadter [1979] presents it too in a more unusual style.

top-down manner — we are guaranteed that such proofs cannot be found. This is the gist of the Hofstadterian upside down causal potency of Gödel sentences. By the same token, the isomorphisms which take place at the level of the mind supposedly inject high-level meanings in the lower-level symbolic network. The high-level self-symbol (which corresponds to agency, consciousness, self-hood and so on) should similarly be regarded as a causal locus with downward causal powers.

5.2. *Brains and formal systems*

It is often strange to bring mathematical or physical results into other areas. For example, we all squirm when we hear Heisenberg's Uncertainty Principle being used in discussions about God or about love or about some other disconnected topic. Gödel's Theorems have received their own share of "use and abuse".[14] In fact, Hofstadter sympathizes with this toward the end of his first book:

> It would be a large mistake to think that what has been worked out with the utmost delicacy in mathematical logic should hold without modification in a completely different area. (1979, p. 696)

We mention this because one common misunderstanding of Hofstadter's work is that he makes use of a psychological variant of Gödel's work. He does not do that and it would be impossible for him to do that. Formal systems and mental systems reveal a handful of crucial distinctions. In Sec. 3, we outlined various aspects of the haziness of symbols and their interconnections and overlaps. There is clear mismatch between the fuzziness that is inherent in our symbolic network which mirrors external reality and the crispness of formal arithmetics which deal with a determinate, precise mathematical reality. This effectively prevents symbolic networks from being capable of serving as theoretical objects of mathematical discourse.

Furthermore, the Gödel sentence is just another arithmetical sentence couched in the formal language of the system, but there is no way to make a robust mentalistic parallel in which we generate the self-symbol out of the symbolic network of a brain in a similar fashion: there is no Diagonal Lemma for mental systems. It is also noteworthy that inside each brain there is a privileged self-symbol, but formal systems admit infinitely many equally-good Gödel-like sentences through which that system supposedly "acquires a self" — none of these is privileged in any way at any instant in time. Lastly, the human self is continuously moulded by brain dynamics, but metamathematical selves are forever unchanging — this is obviously manifested by the obvious fact that we have no counterpart to the death of a self in metamathematics.

Given all of this, it is good news that Hofstadter does not attempt to use a psychologised version of the first incompleteness theorem. That would have truly been a dead end. However, the fact that he does not commit that mistake comes at the cost of watering down the reach of his project to a considerable extent. That is

[14] See Franzén [2005] for an authoritative overview on this topic.

because whilst we understand perfectly well how to construct a Gödel sentence out of an underlying formal system and we know perfectly well why that sentence happens to be undecidable (given our understanding of arithmetization and diagonalization), we are at a total loss as to how we can get a firm grip on the emergence of the self-symbol out of the symbolic network. If the theorems cannot tell us anything in this respect, then the details need to be filled in somehow — otherwise, how exactly does Hofstadter invite us to look at the self-symbol if we are bound to fail to localize it even if we embrace the Hofstadterian "high-level" motto?

All of this sits on top of the more alarming fact that we have no prior grip on the relationship between ordinary symbols such as the DOG-symbol and the lower-level details of our biological neural network. We can only claim to have a very vague idea about how the Hofstadterian brain levels hang together. To say that the emergence of a self-symbol is inevitable in symbolic networks of human brains because their complexity guarantees some sort of internal meta-mirroring is a philosophical credo which cannot really be empirically supported. This is not to say that the hypothesis is wrong: we simply remark that it is practically incapable of being confirmed or falsified given today's resources. However, we judge that there is an air of plausibility to it and we shall be charitable to Hofstadter's suggestion if we do not have principled reasons to reject it.

Thus, there are two difficult questions that emerge naturally at this point in the discussion. On what grounds should we believe that once a brain mirrors the external world through its symbolic network then that must give rise to brain-patterns which will mirror the mirroring? The second question is the usual hard problem of consciousness, but this time taylored to Hofstadter's approach: how can the self-subsystem account for subjective experience? Since in this paper we are more concerned with mental causation, we will not press any harder on the second question, but it is an important one which Hofstadter usually addresses roughly by pointing out that people are stuck in Cartesian thinking traps and that the problem will dissolve for anyone who really understands his picture. Few would be convinced by this explanation, including Chalmers [1996, pp. 30−31], Hofstadter's former doctoral student and the coiner of the phrase "the hard problem of consciousness". Chalmers points out that Hofstadter's analysis is less about subjective experience and more about aspects that have to do with introspection or free will.

Hence, leaving problems of subjective experience aside, we come back to the previous point concerning the lack of details concerning the emergence of the self-symbol due to meta-mirroring. This problem pours cold water on the prospects of yielding fresh perspectives on the relationship between the self and the lower levels of the brain. The reason is simple: we do not really know what to look at to begin with, let alone look at it with new eyes. But Hofstadter does not shy away from this claim and, as we have seen, he maintains that we can indeed appreciate the reality of mental causation after understanding why Gödel sentences have downward causal abilities.

He indeed acknowledges that he's just making an analogy, but there surely must be some standards that we need to impose upon analogies in order to proclaim that we've shed new light on problems that hounded philosophers for centuries. He himself remarks that analogies have to go all the way:

> (I)t is important to realize that if we are being guided by Gödel's proof in making such bold hypotheses, *we must carry the analogy through thoroughly* (...) If our analogy is to hold, then, "emergent" phenomena would become explicable in terms of a relationship between different levels in mental systems. (1979, pp. 708−709, *my emphasis*)

Without an understanding of the syntactic nature of the symbolic network and the *details* of how the self emerges out of it,[15] we have absolutely no way of using our prior understanding of the incompleteness phenomenon in Metamathematics to really be able to confidently say we finally unlocked a fresh understanding of classical mind-body problems. This unfortunate state of affairs already weakens Hofstadter's proposal to a considerable extent, but we temporarily choose to turn a blind eye in the next section.[16] There we ask ourselves: How far can the Gödelian analogy be carried through?

5.3. *Downward causation in metamathematics and in brains*

Did I (viz. *my-self*) decide to sit down in front of my computer and write these hopefully not-too-tedious pages to be peer-reviewed? It certainly *felt* that way to me. However, given the complicated neural computation and symbolic activity which takes place in my skull, does it make sense for me to *really* say that my mind is the causal locus behind all of this? Hofstadter thinks so, and his reasons boil down to an idea that we mentioned a couple of times (and which needs further unpacking): mental causation, just like the incompleteness phenomenon, is an intrinsically high-level phenomenon that cannot be understood, observed, analyzed or grasped at lower-levels.

We will now recapitulate some basic elements from Sec. 5.1. Hofstadter's Gödelian analogy brings into play formal systems (which he acknowledges that they are orders of magnitude less complex than "mental systems") in order to show why certain phenomena can only be made sense of at the appropriate level (and thus at no other level). Each consistent formal system F that can be subjected to the first incompleteness theorem has a formally undecidable sentence $\mathsf{G_F}$, also known as F's Gödel sentence. In order to understand why this sentence and its negation are independent from that formal system, Hofstadter maintains that it would be a complete waste of time to only look at the low-level ingredients of the formal system, namely the rules and axioms, and juggle these in an attempt to eventually derive $\mathsf{G_F}$. No amount of mathematical shuffling will pave the way to the *real reason* of $\mathsf{G_F}$'s undecidability.

[15] Without using imprecise words or phrases that he tends to use such as "symbolic dance".

[16] His project might be weakened, but this does not mean that there is nothing positive to be learnt from his overall Philosophy, as we should see later.

This reason is a *high-level reason* which can only be grasped from a level of analysis where one can undertake syntactic arithmetization and from which various isomorphic relationships can unlock a fresh semantic perspective on G_F. With the Logician's lenses on, we can see an equivalence between G_F and the unprovability predicate of F applied to G_F's name. This is a subtle relationship which effectively prevents G_F from being formally provable: the sentence seems to be imbued with a high-level meaning whereby it apparently self-prescribes the property of being unprovable and, in effect, *guaranteeing this fact*.

It is *because* of that high-level meaning that both G_F and $\neg\mathsf{G}_\mathsf{F}$ lack a formal proof inside F: the metamathematical reading of the sentence seems to exhibit *downward causal* powers over F. When it comes to brains, Hofstadter similarly maintains that no amount at looking at neural computation or symbolic activity can be worthwhile, for mental causation is not a phenomenon that can be grasped (or can even make sense) at these low brain levels. Only by looking at the level of the mind will see upside down causality in action. This fragment from the preface of GEB captures some of the elements discussed so far:

> Something very strange thus emerges from the Gödelian loop: the revelation of causal power of meaning in a rule-bound but meaning-free universe. And this is where my analogy to brains and selves comes back in, suggesting that the twisted loop of *selfhood* trapped inside an inanimate bulb called a "brain" also has causal power — or, put another way, that a mere pattern called "I" can shove around inanimate particles in the brain no less than inanimate particles in the brain can shove around patterns.

In effect, the Gödelian analogy boils down to this: we've seen how downward causation is a meaningful notion in meta-mathematical strange loops. By looking at specific syntactic entities, i.e. formal systems, from a higher vantage point we grasp something that we'd be bound to miss without that high-level perspective. Brains are also syntactic entities yielded through strange loops and, Hofstadter maintains, the level of the mind comes equipped with high-level properties. One of the important high-level properties of self-symbol is mentalistic downward causality.[17]

A first worry one may rightfully have at this point is that of equivocation on the word "causation". Suppose it is indeed correct to say that it is *because* of their high-level meanings that Gödel sentences are undecidable and, in effect, we bypassed the need to search from bottom-up Hilbert-style proofs for that sentence. We thus decide to baptise this state of affairs with the catchy metaphor of "downward causation". Hofstadter then aims to make a parallel with brains and say that maybe brains

[17] In the next subsection, we wonder whether Hofstadter is really entitled to say that the only reason behind the formal undecidability of a Gödel sentence within its mother system is exclusively owed to what the sentence *says* about itself and that there's no other reason which enables us to declare that sentence undecidable. That discussion will be an extra — we now focus on the important question of whether this Gödelian analogy (with the foregoing assumption granted) can really make us understand mental causation differently and look at it with fresh eyes.

exhibit downward causation too. Can it be the case that the *same notion of causation* is used in both contexts?

We should elaborate on this. In the first Gödelian case, we indeed talked about the *reason* behind G_F's undecidability. It is true that, sometimes, the flexibility of language allows us to interchange *reason* with *cause*. But one cannot literally talk of "real" causation in a frozen, static mathematical reality where nothing ever happens or ever happened. The (meta)mathematical relationships and states of affairs are timeless and unchanging. Mathematical reality brims with abstract objects which, virtually by definition, are causally impotent and lack spatio-temporal positions.

On the other hand, the situation could not be more different in brains, where there are a zillion of things happening at every instant in our biological neural network. Our symbolic network (which analogically corresponds to the level of the formal system) is likewise engaged in plenty of symbolic activity. If we look back at the preceding Hofstadterian quote, he is after a novel explanation of how the self can *shove around* inanimate patterns in the brain. But one cannot possibly abet this "dynamic" notion of causation in minds through an analogy with a metamathematical setting in which literally no "shoving around" takes place. In order for an analogy which buttresses upside-down causality to even *hope* to work in the case of mental causation, one first needs to single out an example and a setting where *the same notion of causation that we are interested in* exhibits the curious flip which can only be understood from the high-level perspective.

But, one may briefly interject, the self (on the Hofstadterian view) is an *abstract* object, for it is a pattern stored in the medium of the brain. Doesn't this shared property of abstractness that selves have with mathematical objects license a metamathematical parallel? We maintain that it does not, because even though we talk of abstract objects in both cases, the mathematical properties of abstract objects do not change over time, but the abstract pattern of selfhood undergoes updating process every single moment and eventually disappears after a typical human lifespan. Abstractness is not enough for the metamathematical parallel to go through.

Let us take stock with respect to our points in Sec. 5.2 and this section. In the former, we pointed out that even if we understood downward causation in formal systems, Hofstadter left us no precise instructions about how to look at brains in order to see the causal flip. If one looks at a Gödel sentence as a naked, unchunked, long string of arithmetical symbols, its high-level meaning would *not* be available to us, as we will see in Sec. 5.4. It is only by through carrying out well-known mathematical steps and by using explicit labels and shortcuts that we are able to cognitively grasp the relationships between a Gödel sentence and a system — we would have no means of looking at the metamathematical situation appropriately if we would simply be handed pages and pages of bare arithmetical symbols and nothing more. In the case of the brain we're not even handed the symbols: how exactly are we supposed to look at brains in order to see the high-level causal potency of selfhood? This was the main point from the previous section; the main point of this one was

that the notion of causation does not seem to match in the two instances linked through the analogy. Combining these points makes us extremely skeptical that, after grasping the incompleteness phenomenon in metamathematics, one can suddenly look at the self as a high-level entity with real causal powers.

Even though we will maintain that there is a mismatch in the senses of causality involved, we have to note that Hofstadter has a personal take on the notion of causation which could potentially answer our charge of equivocation, for he believes that there is not that much of a difference between talk of causality and talk of high-level abstract reasons in discussions about global phenomena. Hofstadter [2007, Chap. 3] points out that a deep understanding of causality sometimes transcends physical interactions and "requires the understanding of very large patterns and their abstract relationships and interactions". (p. 41) I will now slightly modify his preferred intuition-pumping example.

Suppose one wonders about a physical hardware aspect which goes on inside a computer when an algorithm (such as Merge-Sort) or some other high-level piece of software is executed. For example, we may we wonder about electrical matters: why doesn't electrical current pass through some specific area of the computer's hardware when the software is running? One may provide a low-level answer in terms of electrical activity in the neighboring area of the physical place we are concerned with, but that would furnish close to zero real understanding. The real insight-affording reason, or *abstract cause*, is revealed by looking at the abstract high-level algorithm and its mathematical properties.

Hofstadter gives an example in this ballpark meant to illustrate why the answer to a question pertaining to the behaviour of a particular computer piece (his computer is built out of domino-pieces — which he calls "the chainium" — and the software that is running is a basic primality checking algorithm) will have to be offered in terms of abstract causes related to mathematical properties:

> (L)et us try to answer the question "Can the primality of 641 really play a causal role in a physical system?" Although 641's primality is obviously not a physical force, the answer nonetheless has to be, "Yes, it does play a causal role, because the most efficient and most insight-affording explanation of the chainium's behaviour depends crucially on that notion". (...) the local, myopic laws of physics take care of everything on their own, but the global *arrangement* of the dominos is what determines what happens, and if you notice (and understand) that arrangement, then an insight-giving shortcut to the answer of the non-falling domino (...) is served to you on a silver platter. On the other hand, if you don't pay attention to that arrangement, then you are doomed to taking the long way around, to understanding things locally and without insight. (p. 41)

Hofstadter invites us to look at the minds of others in a similar global-organizational level. Coming back to the question which kicked off this section: why did *I* (i.e. the selfhood pattern of this paper's author) decide to write this paper?

The insight-affording answer has nothing to do with neural or symbolic events in my brain: it is because I *hoped* that the philosophical community will further engage with Hofstadterian ideas that I find worthwhile. My *thought* is of course grounded in lower-level processes, but it is *only* the high-level description that truly matters. Thus, for Hofstadter, patterns have causal potency and even properties such as primality can be said to play causal roles in physical systems.

We sympathize with all of this. Since we are finite beings, we seek insightful explanations which target the world at our level. We are bound to often talk back and forth between *reasons* and *causes*. Causality is quite an elusive notion once we depart from fundamental levels of reality and, at the level of macro-reality, we indeed often conflate between the two. Suppose you attend a stand-up comedy show and the comedian tells a hilarious joke. What's the correct answer to the question: What caused your laugh? One can surely give a materialistic explanation of what is going on, but the natural answer is: the *funniness* of the joke caused the laugh. This is the way we speak in our high-level language. Hofstadter's project is not to show that mental causation is a miraculous phenomenon or anything of this sort, but only the more modest one of showing that it is a non-trivial high-level phenomenon: a fairytale after all, but one which unlocks understanding. And, as we previously said, Hofstadter is ultimately a hard determinist: he does not think that this sort of mental causation is special enough to unlock free will or anything of this sort. All he points out, correctly we think, is that context of relationships between abstract patterns, talk of causation and talk of high-level reasons gets mixed up in natural language — and that this mix-up is inevitable given the evolution of our cognitive make-up.

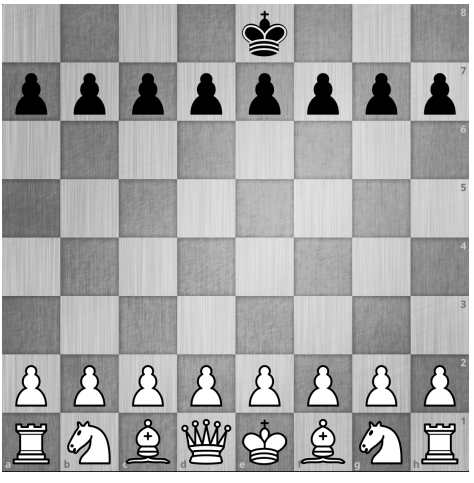
5.4. *Still, why Gödel?*

A pressing question at this stage is the following: Why did Hofstadter feel the need to give so many details about the metamathematics behind the incompleteness theorems? In a way, we mentioned why: we have seen that there are contexts where once a system passes a complexity-threshold, self-reference is automatically achieved and *maybe* that's the fate of the human brain as well. But we've also mentioned that there are strong differences between mental selves and arithmetical selves, weakening the analogy, so Gödel's work can at most show two (related) things: (1) interesting high-level reasons are unavailable from the perspective of the systems and (2) the meaning of Gödel-like sentences exhibits downward causation. We maintain that (2) is not entirely correct based on recent results in Mathematical Logic and that (1) is a simple point which can be presented in a much straightforward fashion outside the arithmetical context.

Let us start with the first one. Hofstadter devotes an enormous amount of pages in both his canonical work and his more recent book on explaining the metamathematics behind Gödel's work in order to show a novel case of arithmetical undecidability and the reasons behind the undecidability of a Gödel sentence. But the *very same* point can be made in a much more straightforward way, as we will do right now. Consider the following familiar formal system: CHESS. This is a syntactic game

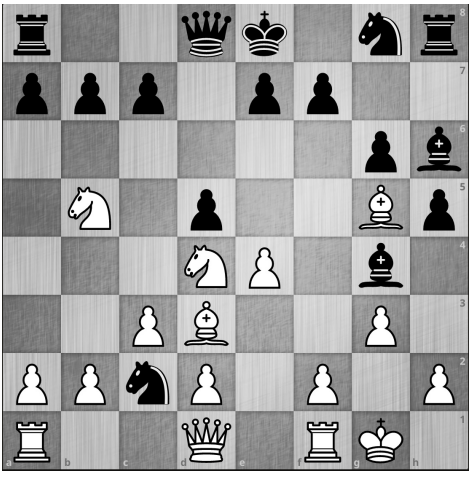
consisting of a 8×8 grid and some pieces called PAWN, ROOK, KNIGHT, BISHOP, QUEEN, KING that play certain structural roles. The CHESS formal system has a handful of syntactic rules for each of the pieces and a single axiom, the starting position of a chessgame. The "theorems" of CHESS are simply chess configurations that are reachable if both players make legal moves.

Question 1: Is the following chess-board configuration a theorem of CHESS?



Answer 1: Even though the above configuration of the pieces looks slightly unusual, it is indeed a theorem—and the insight *can* be obtained at a *low-level* analysis of the board (through a dance of the White and Black Knights).

Question 2: Is the following chess-board configuration a theorem of CHESS?



Hofstadter-style Answer: No, that configuration is *independent* of the CHESS formal system and the reason is *impossible* to be grasped by low-level shuffling of the chess-pieces in search for a derivation. It is only the high-level perspective which yields the true *cause* of independence.

Given CHESS-syntax, the original white bishop which started on square C1 could not be on square G5 unless the original pawns on squares B2 and C2 were moved at some point to make room — but they were not. Can the G5-bishop be an *extra* white bishop acquired later in the game? As we know, the syntactic rules of CHESS allow for pawns to be *promoted* into a main piece when they reach the last square of any file — but no white pawn could have been promoted to a Bishop in our case because White has *all* the eight pawns still on the board.

By taking a high-level syntactic perspective on the board, we revealed the real reason for the syntactic independence of this configuration from its underlying formal system — and this reason is available exclusively at this level.

Remark that we needed no fancy Metamathematics to make the independence claim for this familiar formal system. We made the high-level-reasons syntactic point in the simplest way possible, without missing anything. So what's the purpose of bringing heavy logical concepts into the argument? As we said, it is because of the idea that the self of the system (mental or formal) supposedly exhibits "Downward Causation". In the arithmetical contexts, Downward Causation refers to the impact the high-level meaning of an arithmetical sentence has on its syntactic status (e.g. provable, refutable, independent, undecidable, etc.). Here are Hofstadter's words about the Gödel Sentence of the formal system Principia Mathematica, which he calls KG, that is obtained through the Diagonal Lemma:

> (W)hat is it about KG (or any of its cousins) that makes it not provable? In a word, it is its self-referential meaning: if KG were provable, its loopy meaning would flip around and make it unprovable, and so PM would be inconsistent, which we know it is not (...) [Kurt Gödel's bombshell] revealed the stunning fact that a formula's *hidden meaning* may have a peculiar kind of "downward" causal power, determining the formula's truth or falsity (or its derivability or nonderivability inside PM or any other sufficiently rich axiomatic system). Merely from knowing the formula's meaning, one can infer its [status] (2007, pp. 168−169)

Narrowing in on syntactic properties such as provability, refutability or undecidability, we can formally express Hofstadter's Thesis as follows:

Downward Causal Thesis: If F is a formal system and $\varphi_{\mathsf{F}}(\mathsf{x})$ is a predicate expressing the numerico-syntactic property of encoding a sentence with a certain syntactic status (provable, unprovable or refutable), then any of $\varphi_{\mathsf{F}}(\mathsf{x})$'s fixed points has the syntactic status in question.

So, KG is provable (hence true if Principia Mathematica is sound) simply because KG is a fixed point of the unprovability predicate of Principia Mathematica. But this proposal is quite ambitious. Is it the case that any fixed point of any provability predicate of a certain system has to be a theorem of that system? It wasn't obvious to Leon Henkin who asked the following question in 1952:

> If Σ is any standard formal system adequate for recursive number theory, a formula (having a certain integer q as its Gödel number) can be constructed which expresses the proposition that the formula with Gödel number q is provable in Σ. Is this formula provable or independent in Σ?

The formula that Henkin talks about got to be referred to as *The Henkin Sentence*, i.e. "the sentence which says about itself that it is provable". Hofstadter's preferred arithmetical system, Typographical Number Theory (TNT) can exhibit multiple provability predicates, each of which having Henkin sentences. Through Hofstadterian lenses though, Henkin's question must seem very odd: Isn't it obvious that $\mathsf{H}_{\mathsf{TNT}}$ must be provable simply because it *says* that it is? In other words, isn't $\mathsf{H}_{\mathsf{TNT}}$'s downward causality taking care of the matter?

Well, this is not obviously true, but recent results in Mathematical Logic actually falsify the Downward Causal Thesis. It turns out, borrowing some results from Visser and Picollo, that one can pick three provability predicates for TNT from this bunch with surprising properties:

Theorem 5.1. (The Visser-Picollo Theorems (adapted for Hofstadter's TNT))[18] *There are three provability predicates* $\mathsf{Prov}^1_{\mathsf{TNT}}(\mathsf{x}), \mathsf{Prov}^2_{\mathsf{TNT}}(\mathsf{x})$ *and* $\mathsf{Prov}^3_{\mathsf{TNT}}(\mathsf{x})$ *such that their fixed points* $\mathsf{H}_1, \mathsf{H}_2$ *and* H_3 *are, in turn*:

(1) Formally provable.
(2) Formally *undecidable*.
(3) Formally **refutable**.

Thus, when Hofstadter says that you can directly infer the syntactic status of a sentence such as a Gödel sentence or a Henkin sentence in a straightforward fashion, this is not only controversial, but an actual technical error. His remarks would have predicted that $\mathsf{H}_1, \mathsf{H}_2$ and H_3 are all provable — but only H_1 is. In sum, recent metamathematical results show that the Hofstadterian downward causal thesis, i.e. that the "meaning" of a sentence guarantees its syntactic status, is not a robust one.

This pours some cold water on the Hofstadterian proposal and it is difficult to see what reply he could give at this point. The only possible rejoinder we can think of is something along the following lines: Hofstadter might want to say that he is not interested in provability predicates such as $\mathsf{Prov}^2_{\mathsf{TNT}}(\mathsf{x})$ or $\mathsf{Prov}^3_{\mathsf{TNT}}(\mathsf{x})$ because these predicates do not satisfy the so-called Löb conditions. If a predicate satisfies these three conditions, then it is a theorem that its canonical fixed point H is provable.

[18] Both results appear in Halbach and Visser's [2014] paper. As far as I know, Lavinia Picollo's result is unpublished.

Hofstadter could perhaps maintain that all along he only had predicates which satisfy the Löb conditions in mind. Fair enough, but the ghost of Sec. 5.2 bites back if he chooses to do so. Neither Hofstadter, nor any mathematician can possibly know whether a provability predicate satisfies the Löb conditions *just by looking at the naked predicate*, namely at the extremely long arithmetical string of symbols — which is basically the situation we are find ourselves in when it comes to the brain.[19] When it comes to an expanded provability predicate, nobody which does not have detailed background knowledge of how that predicate was constructed can possibly infer whether the Löb conditions are satisfied. Looking at a gargantuan arithmetical string does not help at all: one needs to chunk that string into components in order to mentally digest it. Even if one can eventually figure out that it is *a* provability predicate, there is no way to tell whether it is subject to the Visser-Picollo theorems.

The point is: without rich understanding of how the relevant syntactic entity was constructed, the "downward causal"-insights simply dissolve and there's no way to bring them back. When it comes to minds, we really have no clue know how to construct a self out of mental symbols. Hofstadter left no instructions: are there any "Löb-like" conditions that minds need to satisfy in order for downward causality to apply?

We have learnt over and over again that we should view the mind from high-levels, but how exactly do we do that? When we look at a Gödel-like sentence to infer its syntactic status (e.g. that it is unprovable), we do that in a very developed mathematico-epistemological background. We cannot do similar "mentalistic" inferences in an *underdeveloped* neuropsychological background.

6. The Current Landscape and Final Thoughts

Because Hofstadter's magnum opus was published over fifty years ago, it is a fairly natural thought to assume that its core mentalistic ideas are outdated.[20] Natural though this thought may be, we maintain that Hofstadter's views are, in a way, quite modern. This section will try to make a short case for that.

To begin with, the Artificial Intelligence literature brims with attempts to use Gödel's Theorems to say something novel about the mind, but virtually all those attempts are carried out in the spirit of Lucas [1961] and Penrose [1989] — they purport to show why it is impossible for the mind to have an algorithmic nature given our ability to grasp Metamathematics. These (very controversial) arguments still see the light of day in Philosophy, Cognitive Science and Artificial Intelligence journals, even though few people actually subscribe to them. For example, a recent stellar discussion of their *logical* shortcomings has been carried out by Koellner [2018a, 2018b].

[19] For an intuitive feel of this, one can check out Hagen von Eitzen's website to see how the fully expanded Gödel sentence $\mathsf{G}_{\mathsf{TNT}}$ constructed out of Hofstadter's TNT system looks like — it fills an entire page. This can be found on his website: http://www.von-eitzen.de/math/tntrep.xml

[20] It may be true that "Gödel, Escher, Bach" was published more than 50 years ago, but his second book on the mind, namely "I am a strange loop", is not that old — it is actually a fairly recent text by scientific standards.

The AI literature also abounds with philosophical discussions which investigate whether such arguments can indeed say something interesting about human minds and/or computational minds. Indeed, a fairly sensible one was carried out by Hofstadter himself [2001] in the preface of the Second Edition of Nagel and Newman's [1958] book.

We shall call the anti-computationalist proposals *Negative Gödelian Proposals*, since they use the theorems in order to showcase the *limitations* of computational accounts of the mental. Given the academic engagement that the negative Gödelian proposals have generated, we thought it's worthwhile to balance them out by putting some of the spotlight on a *Positive* Gödelian Proposal which, despite its resonance, has barely generated detailed discussion in the philosophical literature. As Hofstadter himself points out:

> I must say, I have been surprised and puzzled that the past few years' flurry of books trying to unravel the mysteries of consciousness almost never mentioned anything along these lines. Many of these books' authors have even read and savored *GEB*, yet nowhere is its core thesis echoed. It sometimes feels as if I had shouted a deeply cherished message out into an empty chasm and nobody heard me.

This unfortunate state of affairs is quite puzzling. It is certainly not because contemporary philosophers of mind reject his views. Such a suggestion would be false. For a clear example, Williford [2011] explicitly says that he agrees with almost all of Hofstadter's theoretical claims. Furthermore, the "strange loop" proposal is explicitly endorsed by Dennett [2017], who believes that Hofstadter's work brims with surprising truths. A similar attitude is shared by Mitchell [2019], with whom Hofstadter has developed the CopyCat architecture (whose design incorporates Hofstadter's symbolic tenets). The CopyCat architecture directly influenced the famous LIDA model of cognition (Franklin *et al.* [2016, 2013]) whose biologically inspired design implements both computationally and conceptually the celebrated Global Workspace Theory [Franklin *et al.*, 2012].

In the neuropsychological community, Hofstadter's insights also explicitly motivate the neuropsychological work of Kenneth Williford, Karl Friston and their collaborators [2018]. A handful of Hofstadterian insights can also be found in Metzinger [2009] self-model theory of subjectivity. Moving on from human consciousness to the subjective experience of other life-forms, in his prize-winning investigation of the evolution of consciousness, Godfrey-Smith [2016] makes a handful of philosophical comments that Hofstadter would deeply resonate with.[21]

[21] Dennett [2019] points this out as well in his review of "Other Minds":

> Godfrey-Smith's framework allows him to find valuable contributions to his overall scheme without appearing to be merely a cherry-picker (...) The various *global workspace ideas* fit handsomely (...) *the integrated information* idea gets a constructive gloss, as does Hofstadter's idea of *strange loops* – though without attribution to Hofstadter. (p. 5)

As it can be surely inferred by now, Hofstadter's picture is meant to apply in principle to non-carbon based creatures without problems. In animals, symbols are intimately connected to the central nervous system, but Hofstadter is not particularly concerned with the nature of the pattern-storing media — thus, his picture is meant to apply equally well to silicon-based agents and to machine consciousness in general.

Regarding other philosophical aspects of Hofstadter's project that were presented in this paper, we remark that his ontological relationships with abstract patterns have clearly been inspired by Dennett's work on the topic of "Real Patterns". Dennett [1991] article continues to generate modern discussions (see Millhouse [2021] for a good example). Moreover, Dennett's Pattern-Criterion of Reality is explicitly endorsed by many seminal Philosophy of Science works such as Wallace [2012]. There, Wallace attempts to explain the real emergence of macro-objects out of microphysical levels in the context of the Everett Interpretation of Quantum Mechanics. With respect to Hofstadter's views concerning the abstract causal interplay of patterns, these insights underpin an important section of Deutsch [2011] work, who is particularly fond of Hofstadter's 641-domino example from Sec. 5.3.

In Sec. 4, we presented Hofstadter [2007] more recent views on the relationship between our mature selves and the social aspects which sharpened it throughout our lives. Deep similarities with Hofstadter's explanation of the emergence of the self-symbol through social-feedback loops can be found in Bogdan [2010] who paints a developmental picture of the self in light of sociocultural pressures.

In light of the above, we wholeheartedly believe that Hofstadter's work is replete with fascinating insights. Our purpose in this paper was to investigate the Gödelian aspect of it, which is just a fragment of his broader picture. Even though we think that using heavy Metamathematics is largely unnecessary, this does not beset his overall philosophical project − there are plenty of philosophical lessons to be found in Hofstadter's tomes (which can only be appreciated from the right perspective, otherwise they would be missed).

Acknowledgements

I want to thank Johannes Stern, Max Jones, Tyler Millhouse, Thomas Schindler, Douglas Hofstadter and Melanie Mitchell for their kind help. Some of my thoughts on these issues greatly benefitted from some insightful conversations that I had with Mahmoud Ghanem.

References

Bogdan, R. J. [2010] *Our Own Minds: Sociocultural Grounds for Self-Consciousness.* (MIT Press).

Chalmers, D. J. [1996] The conscious mind: In search of a fundamental theory. Oxford Paperbacks.

Chalmers, D. [2018] The meta-problem of consciousness, *J. Conscious. Stud.* **25**(9−10).

Dennett, D. C. [1991] Real patterns, *J. Philos.* **88**(1), 27−51.

Dennett, D. C. [2017] *From Bacteria to Bach and Back: The Evolution of Minds* (WW Norton & Company).
Dennett, D. C. [2019] Review of Other Minds: The octopus, the sea and the deep origins of consciousness, *Biol. Philos.* **34**(2).
Deutsch, D. [2011] *The Beginning of Infinity: Explanations that Transform the World* (Penguin, UK).
Franzén, T. [2005] *Gödel's Theorem: An Incomplete Guide to its Use and Abuse* (CRC Press).
Franklin, S., Strain, S., Snaider, J., McCall, R. and Faghihi, U. [2012] Global workspace theory, its LIDA model and the underlying neuroscience, *Biol. Inspir. Cognit. Archit.* **1**, 32−43.
Franklin, S., Strain, S., McCall, R. and Baars, B. [2013] Conceptual commitments of the LIDA model of cognition, *J. Artifi. Gen. Intelli.* **4**(2), 1−22.
Franklin, S. *et al.* [2016] A LIDA cognitive model tutorial, *Biol. Inspi. Cognit. Architect.* **16**, 105−130.
Gödel, K. [1931] Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, *Monat. Math. Phys.* **38**(1), 173−198.
Godfrey-Smith, P. [2016] Other Minds: The Octopus and the Evolution of Intelligent Life (William Collins).
Halbach, V. and Visser, A. [2014] Self-reference in Arithmetic I, *Rev. Symbol. Logic* **7**(4), 671−691.
Henkin, L. [1952] A problem concerning provability, *J. Symbol. Logic* **17**, 160.
Hofstadter, D. R. and Dennett, D. C. [1981] *The Mind's I* (Basic Books).
Hofstadter, D. R. [1979] *Gödel, Escher, Bach* (Basic Books).
Hofstadter, D. R. [1985] *Metamagical Themas* (Basic Books).
Hofstadter, D. R. and Mitchell, M. [1994] The Copycat project: A model of mental fluidity and analogy-making.
Hofstadter, D. R. [2007] *I am a Strange Loop* (Basic Books).
Koellner, P. [2018a] On the question of whether the mind can be mechanized, I: from Gödel to Penrose, *J. Philos.* **115**(7), 337−360.
Koellner, P. [2018b] On the question of whether the mind can be mechanized, II: Penrose's new argument, *J. Philos.* **115**(9), 453−484.
Lucas, J. R. [1961] Minds, machines and Gödel, *Philosophy*, 112−127.
Metzinger, T. [2009] *The Ego Tunnel: The Science of the Mind and the Myth of the Self* (Basic Books).
Millhouse, T. [2021] Compressibility and the reality of patterns, *Philos. Sci.* **88**(1), 22−43.
Mitchell, M. [2019] *Artificial Intelligence: A Guide for Thinking Humans* (Penguin, UK).
Murphy, G. [2002] *The Big Book of Concepts* (MIT Press).
Nagel, E. and Newman, J. [1958] *Godel's Proof*, Rev. edn. (New York University Press, New York/London, 2001), edited and with a new foreword by Douglas R. Hofstadter.
Penrose, R. [1989] *The Emperor's New Mind* (Oxford University Press).
Sellars, W. [1963] Science, perception and reality.
Sperry, R. [1966] Mind, brain, and humanist values, *Bull. Atom. Sci.* **22**(7), 2−6.
Wallace, D. [2012] *The Emergent Multiverse: Quantum Theory According to the Everett Interpretation* (Oxford University Press).
Williford, K. [2011] I am a strange loop, *Philos. Psychol.* 861−865.