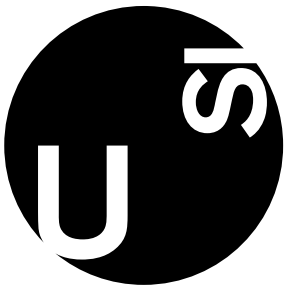

Università
della
Svizzera
italiana

Faculty
of
Informatics

Bachelor Thesis

June 20, 2024

Computational Free Will

A strange loop in the Quantum Turing Machine

Sofia d'Atri

Abstract

An attempt at defining free will, evolving from traditional philosophical definitions to a computational perspective, drawing inspiration from Douglas Hofstadter's "I am a Strange Loop" and Scott Aaronson's "The Ghost in the Quantum Turing Machine". This thesis defines free will in four different ways, discussing objections to its existence and exploring the implications of brain uploading, consciousness in machines and quantum mechanic's role in freedom.

Advisor
Prof. Stefan Wolf

Advisor's approval (Prof. Stefan Wolf): Date:

Contents

1 Introduction

1.1 The problem of free will in technology

Can a machine possess true autonomy? *Are we even truly free?*

To understand the concept of free will, one must piece together elements from seemingly different puzzles, only to discover that they actually fit together. These pieces come from philosophy, neuroscience, computational theories, and quantum mechanics. This thesis attempts to create a large mosaic of these domains and draw a new perspective on the problem of free will, taking it from its traditional philosophical roots to a modern computational perspective. Heavily inspired by the works of Douglas Hofstadter and Scott Aaronson, I aim to bridge the abstract sphere of consciousness with the material mechanics of computation.

The challenges of understanding consciousness and free will are at the core of this investigation, raising questions that challenge the very essence of human experience.

- What are the implications of brain uploading?
- Can consciousness be replicated within machines?
- How does quantum mechanics influence our understanding of freedom?

While these might not seem relevant problems at the moment, as Aaronson notes, they might be in the near future.

> *If any of these technologies - brain-uploading, teleportation, the Newcomb predictor, etc. - were actually realized, then all sorts of "woolly metaphysical questions" about personal identity and free will would start to have practical consequences.* [1, p. 18]

These questions not only push the boundaries of philosophy and science but also urge us to reconsider the essence of what it means to choose freely.

In this thesis, I will define free will by examining its traditional definitions and the nature of consciousness, while analyzing these concepts from a computational standpoint, to finally apply these findings to the scope of technology. To derive these findings, I focused on studying philosophical literature, particularly influenced by Leibniz's philosophy, and engaged in rigorous discussions with my advisor and colleagues.

1.2 Literature review

Understanding free will has been a focus for philosophers and scientists for centuries. Leibniz's ideas are particularly important, and will lay the foundations for the interpretation of the nature of free will and consciousness. His exchanges with Clarke are useful to get a better understanding of the debates on determinism and freedom of choice [15, 27].

More recently, Douglas Hofstadter's book "I Am a Strange Loop" [11] and Scott Aaronson's "The Ghost in the Quantum Turing Machine" [1] offer new perspectives by combining philosophical ideas with computer science. Hofstadter explores how self-awareness works by using the idea of self-referential systems, which he calls strange loops, while Aaronson looks at how quantum mechanics might influence our understanding of free will. David Chalmers, in "Reality+: Virtual Worlds and the Problems of Philosophy" [4], extends the problem of consciousness to virtual worlds, machines and simulations.

This thesis builds on these key works by combining traditional philosophical ideas with quantum mechanics. Through this synthesis, I hope to shed some light on one of humanity's most haunting questions: can a being truly possess the freedom to choose, or are we all, in the end, mere products of mechanistic processes? At the intersection of philosophy and technology, an exciting journey awaits us.

1.3 Acknowledgments

I am deeply grateful to Stefan for his guidance, patience, and expertise as my advisor throughout this journey. His feedback and encouragement have been invaluable in shaping this thesis.
I extend heartfelt thanks to Jeferson, Eduardo and Fanny for the lively discussions we've shared alongside Stefan. Your perspectives and insights have played a crucial role in shaping the ideas presented in this thesis. I also want to express gratitude to all my friends who joined these discussions and contributed to the exchange of ideas.
To everyone mentioned above, thank you for your support and encouragement.

2 A note on definitions

The concept of *free will* will be defined in different ways throughout the thesis. When I mention just *free will*, I mean the generic idea that one is free to do and think what they want, without coercion. However, given the many shades and definitions of free will, I will often write it with the expression *"free will as freedom of x"*, to differentiate between the different kinds.

3 Overview

3.1 Types of free will

In the following pages I will define four types of freedoms. These definitions are closely related, but different arguments can be made against each of them.

1. Freedom from necessity
2. Freedom of choice
3. Freedom of indifference
4. Freedom from predictors

The first three are more traditional definitions of free will. **Freedom from necessity** is the freedom from external, necessary causes. There is no physical necessity for the will to be as it is, so it is free to be as it is not. It is included in the idea of **freedom of choice**, which adds to this that one is free if their choices are not coerced and made only by themselves. A special case of freedom of choice is **freedom of indifference**, which states that, if all our choices are not free but guided by preferences and desires, a choice in a state of true indifference is truly free.

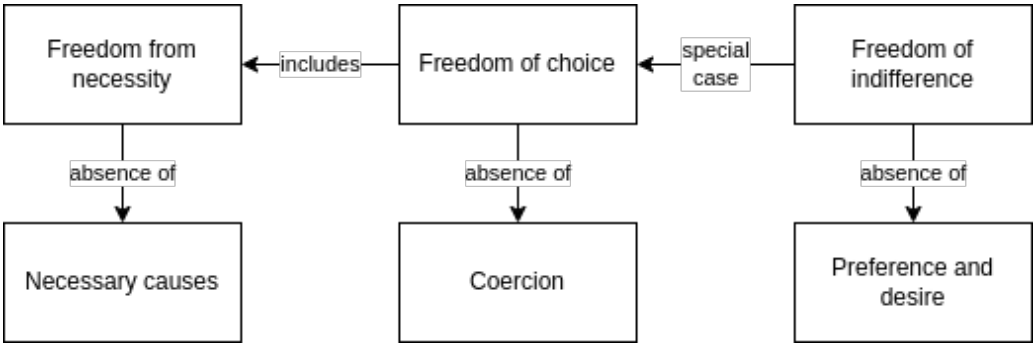

Figure 1. Relationship between the first three kinds of freedom.

Lastly, we move to **freedom from predictors**, an idea originated by Scott Aaronson [1]. He claims that if we can be fully predicted, we likely are not free but, if for some reason it is impossible to predict human behaviour with arbitrarily advanced technology, we *might be* free. This introduces a scientific approach to the free will debate.

3.2 The arguments against free will

In this work, I will only present three arguments against free will, but I am sure that many more exist and can be created. The arguments are the following:

1. The Consequence Argument [12]
2. The Standard Argument [8]
3. The Machine Argument

These arguments against free will are of philosophical nature. While I mention them somewhere later on, I do not analyze neuroscientific arguments, as it is not my field and state-of-the-art neuroscience is outside the scope of this thesis. Moreover, the common consensus is that free will can't be explained quite yet even by neuroscientists.

4 Freedom from Necessity

When you drop a ball, you're confident it will hit the ground. If it's made of the right material, it might also bounce a bit. Knowing the ball's weight, material, drop height, and factors like air and floor friction, you can predict with some precision exactly how the ball will behave. This predictability comes from classical physics, where every action has a **necessary cause**.

In contrast to this mechanical certainty, the concept of **free will** introduces the idea that human actions are not bound by necessity. Unlike the predictable fall of a ball, by accepting free will we accept that our choices are not predetermined by prior states.

> "The term 'free will' [...] has traditionally meant [...] lack of necessity in human will, so that 'the will is free' meant 'the will does not have to be such as it is' " [5].

The will does not have to follow physical rules. It is subject to arbitrary changes by its holder.
Freedom from necessity has been treated only briefly here, but it is an important concept and it will be the basis for the rest of the discussion in this thesis.

4.1 Determinism

Determinism is the idea that everything, including human decisions and actions, is the necessary result of antecedent causes. It also includes the idea that the entirety of the past and the future are completely determined since the origin of the universe [6]. This is the main objection to *freedom from necessity*, as its own premise is that the will is in fact not free from necessity.

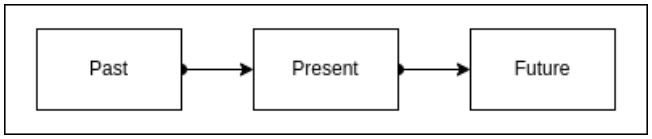

Figure 2. According to determinism, events from the past are the necessary causes for events in the present and in the future. Image inspired from the one on [22].

This is the case of Hobbes's and Spinoza's necessitarian view that everything happens by metaphysical necessity [27] or Peter van Inwagen's **Consequence Argument** [12].
According to the Consequence Argument [1, p. 23], if determinism holds, our choices today are determined by the state of the universe at a remote point in the past, beyond our ability to influence, leading to the conclusion that under determinism, our choices are not within our control, hence negating the existence of free will.
As for metaphysical necessity, it won't be discussed here, since Leibniz and Clarke's arguments against it [27] are the basis of the next chapter: free will as freedom of choice rather than freedom from necessity.

5 Freedom of Choice

"Action there can be none, without Liberty", Clarke [27, p. 80]

An action, whether taken or abstained from, is invariably the outcome of a decision-making process. Apart from involuntary actions which we have no control over - such as the pumping of the heart, the retraction of a hand from a fire, or sneezing - every deliberate action or inaction is rooted in a **causal trigger**. Be it the choice to traverse a particular path, the selection of an item, or deciding to speak (whether you think carefully before speaking, is another issue): each of those is the manifestation of a preceding decision. Conversely, deciding not to engage in a specific action or activity, likewise constitutes a decision, albeit one of non-action. While the majority of decisions we make occur in split-second moments, often without our conscious awareness, such choices, whether affirmative or negatory, originate from a cognitive process in which we weigh different factors and ultimately make a decision.

The ability to make **choices without coercion** is what we will refer from now on as *freedom of choice*. As long as no one is forcing you to do something, you are free to do anything you please.

5.1 Exchange between Leibniz and Clarke

The following discussion is mainly based on the exchange of letters between Leibniz and Clarke and the analysis done by Vailati [27]. Throughout this exchange, they have disagreed on almost everything, but they seem to be somewhat similar at least in their notion of free will, both defining it as a form of *freedom of choice*.

For Clarke, true liberty means having an **internal principle of action**, giving beings the ability to initiate movement or action from within, rather than being passive recipients of external causes. He criticizes necessity by declaring absurd the idea of an infinite chain of causes.

> [It] is a plain Contradiction [...] to suppose an infinite Series of dependent Effects none of which are Necessary in Themselves or Self-Existent; [...] there must be in the Universe some Being, whose Existence is founded in the Necessity of its Own Nature; and which, being acted upon by Nothing beyond itself, must of Necessity have in itself a Principle of Acting, or Power of beginning Motion, which is the Idea of Liberty. [27, p. 80-81]

Instead, Leibniz offers a middle ground with his compatibilist perspective. He agrees that every event has a cause, but he also believes in freedom. For Leibniz, free will has two parts: control over one's **passions**, allowing one to act as one should, and the **absence of necessity** in decision-making, that is our choices are not absolutely determined by prior causes.

In general, both agree that freedom of choice exists when decisions are made without external pressure or emotional influence. The first part is straightforward: if someone makes me do something, I am not free. Conversely, if I'm acting on my own, I'm not forced. So, if forced means not free, then not forced means free. This is generally accepted in everyday life, outside the philosophical debate on free will.

However, the influence of emotions and desires (or passions) is a more subtle point and easier to challenge when arguing that free will doesn't exist. How passions influence our decisions will be discussed in detail later. For now, we can just assume that a person has free will if they are *in theory* able to make choices with control over their own emotions.

5.2 The Hedge Maze of Life

In *I am a Strange Loop*, Hofstadter writes that we will never have complete freedom as most things in life are out of our control. In this sense, we are free, but within the walls of a hedge maze.

> **The Hedge Maze of Life** *Sometimes our desires bang up against obstacles. Somebody else drank that last soft drink in the refrigerator; the formerly all-night grocery store now closes at midnight; my friend's car has a flat tire; the dog ate my homework; the plane just pulled out of the gate*

thirty seconds ago; the flight has been canceled because of a snowstorm in Saskatoon; we're having computer troubles and we can't seem to make PowerPoint work in here; I left my wallet in my other pair of pants; you misread the final deadline; the reviewer was someone who hates us; she didn't hear about the job until too late; the runner in the next lane is faster than I am; and so on and so forth. In such cases our will alone, though it pushes us, does not get us what we want. It pushes us in a certain direction, but we are maneuvering inside a hedge maze whose available paths were dictated by the rest of the world, not by our wants. And so we move willy-nilly, but not freewilly-nilly, inside the maze. A combination of pressures, some internal and some external, collectively dictates our pathway in this crazy hedge maze called "life". [11, p. 339]

This idea can only apply to *free will as freedom of action*, while in principle it shouldn't apply to *freedom of choice*, as choice does not have an action as a necessary consequence: one could make the decision to think about something or to refrain from some action. So, in theory, it does not limit *freedom of choice*, but puts boundaries on *freedom of action* in the real world.

5.3 The Standard Argument

The **Standard Argument** [8] is another argument that uses determinism against free will. It is discussed here - and not in the previous chapter - since the focus, rather than being on necessity, is on the control we have over our will and decisions. It goes like this:

1. If determinism is the case, the will is not free.
2. If indeterminism and real chance exist, out will would not be in our control.

There are some interesting **objections** against this argument, but I will just highlight two, both directly quoted from [8].

1. Determinism is not "true". If anything physical is "true", it is indeterminism.
2. Randomness in some microscopic quantum events is indeed chance. But microscopic chance does little to affect adequate macroscopic determinism.

With this, we will move on from the idea that everything is predetermined at the beginning of the universe. However, we will get back to randomness and microscopic events later on and show how they actually help us defend free will.

5.4 The Machine Argument

Rather than determination since the beginning of time, some have argued for what I will call the **Machine Argument**: humans are machines that react in predictable ways to stimuli, thus their choices are not free. A variation of this argument posits that a person makes decisions solely based on their life experiences; hence, at every moment, any action they take is the only one they could have ever taken. The main flaw in this argument is that it is based on a posteriori observation: it is easy to say that a person could have acted only in a specific way after it has happened. Moreover, it opens a Pandora's box of dangerous implications, such as the idea holding people in a "quarantine prison" before they can commit a crime they have been predicted to commit based on their life story.
Let's break the Machine Argument in two parts and analyze them one by one. Since the argument supposes that, given a *person* and their *life experiences*, their actions are predictable, we can call these components the **Biological** one and the **Historical** one. The Biological component would be a person's brain, together with all the structures that allow for perception and elaboration of stimuli, in their unique way. The Historical component is the suite of all life experiences of a person, that shape their personality.
In the following paragraphs, I will argue that *freedom from the biological component* is not relevant, following Hofstadter 's metaphor of the *careenium*, nor desirable, and that the Historical Component is not enough to negate free will on its own.

In the next chapter, I will instead start with the assumption that the Machine Argument is true: all decisions we make are either due to some biologic or historic desire. Even in this case, as we will see, it is possible to define free will.

5.5 Biological freedom

The assumption is that, since the brain is composed of neurons that function mechanically, any decision we think we make freely is actually just the result of these neuronal processes. Therefore, we are not truly free but rather biological, adaptive machines, responding to stimuli in predictable ways.

Before neuroscience, one could also try to define freedom in a molecular sense, or atomically or even in a subatomic way. But it sounds absurd to delegate the responsibility for all actions of humankind (and animals too!) to simple particles. Sure, at the end of the day we depend on our neurons to function. However, the *human experience as seen from the inside* feels richer and deeper than that of a mechanical being. Free will, freedom and consciousness are concepts which lie on an abstraction plane higher than that which molecules can explain.

To understand better what I mean with the concept of higher abstraction, here are a couple passages from *I am a Strange Loop*, in which Hofstadter uses the metaphor of "simmballs" to explain the idea of abstraction in the mind.

The Careenium *Imagine an elaborate frictionless pool table with [...] myriads of extremely tiny marbles, called "sims" (an acronym for "small interacting marbles"). These sims bash into each other and also bounce off the walls, careening about rather wildly in their perfectly flat world — and since it is frictionless, they just keep on careening and careening, never stopping. [...] The sims are also magnetic (so let's switch to "simms", with the extra "m" for "magnetic"), and when they hit each other at lowish velocities, they can stick together to form clusters, which I hope you will pardon me for calling "simmballs". A simmball consists of a very large number of simms, and on its periphery it frequently loses a few simms while gaining others.*
There are thus two extremely different types of denizen of this system: tiny, light, zipping simms, and giant, ponderous, nearly-immobile simmballs. The dynamics taking place on this pool table — hereinafter called the "careenium" — thus involves simms crashing into each other and also into simmballs.[11, p. 45]

Simmballism *We can posit that one particular simmball always reacts in some standard fashion to breezes, another to sharp blows, and so forth. Without going into details, we can even posit that the configurations of simmballs reflect the history of the impinging outer-world events. In short, for someone who looked at the simmballs and knew how to read their configuration, the simmballs would be* ***symbolic****, in the sense of encoding events.*[11, p. 46]

Taking a Higher-level View of the Careenium *We spatially back away or zoom out, thus rendering simms too small to be seen, and so the simmballs alone necessarily become our focus of attention. Now we see a completely different type of dynamics on the table. Instead of seeing simms bashing into what look like large stationary blobs, we realize that these blobs are not stationary at all but have a lively life of their own, moving back and forth across the table and interacting with each other, as if there were nothing else on the table but them. Of course we know that deep down, this is all happening thanks to the teeny-weeny simms' bashing-about, but we cannot see the simms any more. In our new way of seeing things, their frenetic careening-about on the table forms nothing but a stationary gray background.*[11, p. 49]

Who Shoves Whom Around inside the Careenium? *In one view, the meaningless tiny simms are the primary entities. [...] In this view the simmballs are not even recognized as separate entities, since anything we might say about their actions is just a shorthand way of talking about what simms do. From this perspective, there are no simmballs, no symbols, no ideas, no thoughts going on — just a great deal of tumultuous, pointless careening-about of tiny, shiny, magnetic spheres.*
In the other view, [...] the interest resides solely in the simmballs, which give every appearance of richly

interacting with each other. One sees groups of simmballs triggering other simmballs in a kind of "logic" that has nothing to do with the soup churning around them, except in the rather pedestrian sense that the simmballs derive their energy from that omnipresent soup. Indeed, the simmballs' logic, not surprisingly, has to do with the concepts that the simmballs symbolize. [11, p. 49]

The Dance of the Simmballs *From our higher-level macroscopic vantage point as we hover above the table, we can see ideas giving rise to other ideas, we can see one symbolic event reminding the system of another symbolic event. [...] The simms are still there, to be sure, but they are simply serving the simmballs' dance, allowing it to happen.*[11, p. 50]

Hofstadter uses the metaphor of "simmballs" to explain the relationship between neurons and mental states. Imagine a frictionless billiards table with lots of tiny magnetic marbles, or "simms," bouncing around. These simms stick together to form clusters called "simmballs." The table's walls react to outside forces, affecting the simms and simmballs inside. This setup represents how neurons (simms) form patterns (simmballs) that encode information about the environment.

His point is to show that we can view the brain in two ways: by looking at the tiny neurons following physical laws (simms) or by looking at the larger patterns they form (simmballs) that represent thoughts and ideas. If we speed up the motion and zoom out, the simmballs become the main focus, symbolizing how our thoughts and ideas interact.

The main takeaway from this is that it is hard (I would say impossible in everyday life) to describe thoughts, ideas and emotions solely in terms of neuronal interactions. We build ideas on top of ideas, construct conversations based on thoughts. Our happiness may be just hormones being released, but to us it is seeing someone dear, petting a cute fluffy dog, eating our favorite food, frolicking in the fields picking up flowers. Hormones get released in the "low level" background while we experience the "high level" world. I am not claiming we are separate from our Biological component; rather, it shouldn't be the only basis to dismiss free will. Disconnecting from our biological needs would be disastrous. Our bodies constantly signal what they need - food, rest, healing, movement, or relief from overstimulation. We have some freedom in responding to these signals: you might eat now or later, or endure a stressful meeting even when your body wants out. Ignoring these signals leads to consequences like hunger or headaches. Imagine being completely independent from these bodily signals; you wouldn't feel your body's needs and might neglect essential actions like eating or resting until it's too late, causing serious harm.

In conclusion, I claim here that a *freedom from biology* would be undesirable, but that our dependence on it is not enough to argue against free will. Whether our actions are exclusively governed by our nervous system, is a question for neuroscientists to answer. And so far, they don't have a certain answer yet.

5.6 Neuroscience interlude

In the introduction I mentioned neuroscience being outside the scope of this thesis, but it is worth mentioning two notable experiments and their results here. I will not discuss them in detail, and I am sure more experiments have been conducted that could prove, disprove or add new results to these findings. The relationship between brain activity and free will is fascinating but complex, and these experiments are just the tip of the iceberg, showing that while neuroscience helps describe brain phenomena, it alone cannot fully unravel the mysteries of human decision-making.

5.6.1 Libet's experiment

In 1983, neuroscientist Benjamin Libet conducted a famous experiment where participants were asked to flick their wrist while noting the time they decided to do so [7, 16]. Brain scans showed that the brain began the action before the participants reported making the decision (see Figure 3), suggesting that the brain might decide *before* the conscious mind is aware.

Libet followed up with the idea of a "veto" mechanism: when participants were asked to stopped mid-move when hearing a buzzer, the brain activity was shown to stop without a conscious decision, hinting at an internal "free won't", or veto, system.

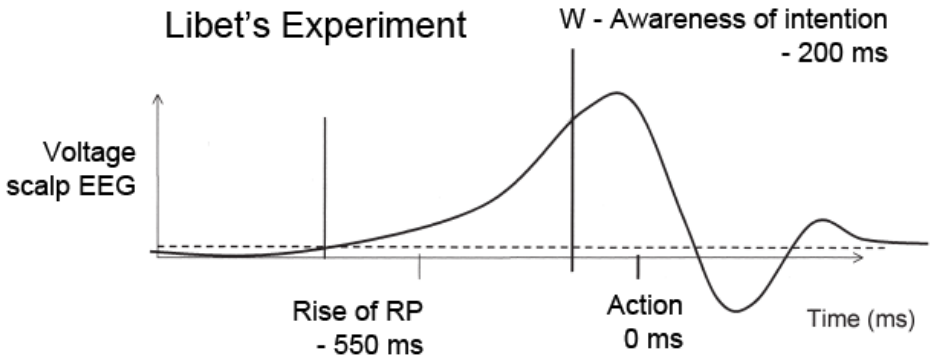

Figure 3. Libet experiment. Image from [7]

This led to debates on whether our choices are truly free or just rationalizations of unconscious processes. Some critics have argued that remembering the clock's time is prone to introducing errors and noise in the measurements, and that wrist flicks aren't the same as more complex, life-altering decisions. Successive studies, like those by Aaron Schurger and later researchers, aimed to replicate and challenge Libet's findings, with mixed results [20].

5.6.2 Mystery of the Mind

Wilder Penfield's work further complicates the picture. Penfield, an American-Canadian neurosurgeon, conducted thousands of brain operations on conscious patients and discovered that while he could stimulate parts of the brain to induce movements, sensations, and memories, he couldn't alter a patient's free will or moral beliefs [17]. For instance, he could make a patient move their arm but couldn't make them believe that $2+2=5$. Similarly, seizures, which are intense bursts of electrical activity in the brain, never seemed to affect a person's core beliefs or free will.

These experiments suggests that while the brain controls many functions, the essence of free will might lie beyond its direct influence. As research progresses, so will our understanding of free will, but for now, scientists don't have a definitive answer yet.

5.7 Historical freedom

When someone commits an unexpected and violent act, people often dissect the aggressor's life to try and find motivations. "Of course they did that, given their past," they say.
People who commit similar acts are grouped together, reinforcing the idea that a certain past dictates a certain future. But this reasoning only works after the fact. We connect the dots in hindsight, claiming it all makes sense. Or worse, we say, "there were no signs this could happen".
Hindsight is always 20/20, but it doesn't prove that someone with a particular history would necessarily act a certain way. Two people with similar life experiences can end up completely different. Until we can travel back in time within our universe, we won't be able to prove or disprove whether we have *freedom from historical events*.
While the Biological component can be scientifically explained and, after enough research will have been done, could be a valid argument against free will, the Historical one can't be formalized due to its nature. In conclusion, while the Historical component influences us, it's not enough to argue against free will.

In the last chapter I will discuss the implications of brain cloning and simulation, which is closely related to the History component.

6 Freedom of Indifference

"In things absolutely indifferent, there is no [Foundation for] Choice; and consequently no Election, nor Will; since Choice must be founded on some Reason, or Principle." Leibniz, in his 4th letter to Clarke [15]

Can a reasonable agent make a **choice without a clear preference** for either option? Since the agent is reasonable, they should have a reason for their choice. But in a state of equilibrium, where one alternative is the same as the other and the agent has no preference for either, it seems that the agent should be unable to make a choice. Thus we can say that *no reasonable choice can be made without a preference*.

6.1 Identity of Indiscernibles

A little note should be made here. Stating that two different alternatives are the same does not mean that there exists no difference between them. The problem of the *identity of indiscernibles*, central in Leibniz's philosophy, is not the reason for the inability to choose.

Leibniz's law *If x is identical to y, then every property of x is a property of y, and vice versa.* [25]

In the case of two qualitatively identical items, their distinctive property would be their position in space. However, this is not the main focus in our discussion. When saying that *one alternative is the same as the other*, it is intended that, while the two options are **different** in the sense that they don't share all their properties, they are **identical** in the most relevant sense. The alternatives possess differences that are either irrelevant concerning their desirability or unknown to the agent: all of the known reasons for desiring either alternative are equivalent.

6.2 Buridan's Ass

An example for such situation would be the paradox of Buridan's ass, a thought experiment wrongly attributed to Buridan but existing long before him: we find some of the first formulations dating back to Anaximander, Aristotle and Ghazali. [19]

This though experiment places a hungry donkey between two bundles of hay which are absolutely identical, at equal distance, and of equal desirability. The animal has no reason for favoring one or the other, but it must choose one, or else it will starve to death. The assumption is that the creature is a reasonable agent, thus it will prefer eating either bundle rather than not eating any. At the same time, given that there exist no reason for picking one over the other, if we follow the principle that *no reasonable choice can be made without a preference*, the donkey will fail to make a choice and starve.

This deadlock appears absurd and impossible, thus there must be something that allows a choice to be made in the absence of preference. Many philosophers gave their resolution to this problem, by providing different explanations of what could be the force that allows a choice without preference. Some used this freedom of indifference to identify free will, while others used it as a demonstration of free will's inexistence. As mentioned previously, while this paradox is wrongly attributed to Buridan, he was familiar with it and, in fact, gave the example of a dog starving between two equal portions of food [26]. Buridan answers the problem of choice without preference by holding that the **will is subject to reason**. He holds that reason judges the options and decides which is the greater good. In turn, the will can only opt for the greater good.

While he limits the freedom in the context of indifference, he gives the will the characteristic of a *facultas suspensiva*, the ability to halt the decision process and allow reason to further examinate the options and, possibly, deem the greater good a different option than the one that was originally chosen.

6.3 Minute Perceptions

Claiming that the freedom of indifference is the manifestation of free will is not enough: assuming that it implies free will's existence, how can we be sure that we're capable of truly indifferent choices? When offered two coins of the same value, a person might choose the one on the right because they're right-handed. This is an example of influence from *petites perceptions*, here translated as *minute perceptions*, those perceptions according to Leibniz which remain below the threshold of conscious experience, and thus go unnoticed in the decision-making process.

While Leibniz argues in favour of free will, as discussed earlier, he is against the idea of *free will as freedom of indifference* [27]: he believes that a person is influenced by some factors that are unknown to them. A precursor in what would later be known as Freud's theory of the unconscious [13, 24], Leibniz's idea of unconscious mental representations was a radical and new idea, though it wasn't as widespread in his times [13]. These minute perceptions, although not immediately apparent to us, subtly shape our thoughts and actions. They accumulate and influence our decisions without us being directly aware of their presence. Similarly for Freud, these hidden mental states explain human behaviour and determine it [24]. Thus in the case of a choice in a state of indifference, there must be, according to Leibniz, some minute perceptions that guide us which we aren't fully aware of. Additionally, Leibniz did not claim that we cannot in principle become aware at a later time that we experienced a minute perception in the past [27, p. 93], so it is entirely possible that we are not aware of the motives behind an action we thought we performed in indifference until later.

Furthermore, the brain itself can be seen as a chaotic system, where small initial differences in neural activity can lead to vastly different outcomes. This chaotic nature means that our decisions are highly sensitive to minute, seemingly insignificant influences. From this, it follows that whenever we make a choice, we are influenced by hidden mental processes beyond our awareness and control. So, according to this perspective, free will would be more limited.

6.4 Randomness in Indifference

Are choices made in a state of indifference irrational, or a-rational even? If no reason is involved in preferring one item over another, then the rational choice would be to not choose at all.
Does this mean that choices made in a state of indifference are random? From an outside perspective, they seem to be. For example, if I offer you two identical coins and you must pick one, you'll just grab one because they're the same in value and style. An external observer would think your choice was random, unaware of the mechanisms in your brain. And you would probably agree: you just had to get your change back, so it doesn't really matter what coin it is as long as you get the right value back. A few questions thus arise:

- Are humans acting randomly or irrationally in such cases?
- If we act randomly, do we have a built-in randomizer? [19, p. 25]
- Are these choices truly random, or are they influenced by the subconscious, as Leibniz and Freud proposed?
- Can random choices be made even with the influence of the subconscious?

Lastly, how does randomness relate to the concept of free will?

7 Freedom from Predictors

7.1 Unpredictability

Some people argue that with enough knowledge, human behavior can be quite predictable. This predictability would be the foundation for marketing, the effectiveness of advertisements, and the ability to conduct statistical analyses on population interests.

Additionally, it can be easily argued that predictability and freedom don't imply each other.

- Assuming a person has free will, they could choose to live a very repetitive life, thus becoming predictable while being free.
- Similarly, a system being unpredictable doesn't imply the presence of freedom: chaotic systems are deterministic after all.

So is unpredictability really the key to freedom?

7.2 The Ghost in the Quantum Turing Machine

So far, unpredictability hasn't been used as an argument in the free will debate. Yet.
Enter Scott Aaronson, who described a new way of looking at the problem of free will starting exactly from unpredictability, and quantum mechanics. His interest is whether quantum mechanics imposes interesting limits on *an external agent's ability to scan, copy, and predict human brains* and other complex biological systems. His goal isn't to resolve the free will debate but to remove some small solvable piece from the big question, re-frame it in a physical way, and offer a plausible and testable picture, stripping away ethical, metaphysical, or linguistic considerations and focusing solely on physical predictability.
In other words, he replaces "*the question of whether humans have free will, [with] the question of how accurately their choices can be predicted, in principle, by external agents compatible with the laws of physics*" [1, p. 13]. He then makes the following claims:

1. If humans are found to be predictable in the relevant sense, it would support the idea that free will is an illusion.
2. If fundamental reason were discovered that make it impossible to predict humans, then it would support the notion that free will is real.

Here, I baptize this kind of free will as *freedom from predictors*. One is found to be free if they cannot be predicted, not even in principle, by a *Predictor*, an agent that would make a copy of a human's brain and predict their every move.

7.3 Knightian Uncertainty

Frank Knight was an economist who shook up the field with his 1921 book "Risk, Uncertainty, and Profit", where he made a key distinction between risk and uncertainty. According to Knight, risk involves situations where we can't predict the outcome but can measure the odds. Uncertainty, however, is when we can't even figure out the odds because we lack crucial information: some future events are fundamentally unpredictable and beyond our ability to quantify.

7.4 Freebits

Here is an overview of the ideas presented in Aaronson's essay.

He introduces freebits: "*due to Knightian uncertainty about the universe's initial quantum state, at least some of the qubits found in nature are regarded as freebits*" [1, p. 36] which make "*predicting certain future events - possibly including some human decisions - physically impossible, even probabilistically*" [1, p. 36]. Freebits are qubits because they can't be measured without violating the no-cloning theorem, and their

causal history cannot be traced back to any physical process that generated them according to a known probabilistic ensemble.

He then differentiates between “macrofacts” and “microfacts”. Macrofacts are about decohered, classical, macroscopic properties of a physical system that can be learned without disturbing the system. Microfacts, on the other hand, pertain to undecohered quantum states that can’t be known without potentially altering the state if measured in the wrong basis. Freebits fall under microfacts, as their quantum state always involves Knightian uncertainty, making the prediction of certain events, possibly including human decisions, physically impossible even with advanced future technology.

In in the world of microfacts, there is no such thing as causality. Microfacts don’t have a clear causal direction. Due to unitarity, when a microfact f “causes” another microfact f' the temporal order of f and f' irrelevant.

$$U|\psi\rangle = |\phi\rangle$$

then

$$U|\phi\rangle = |\psi\rangle$$

Causality, tied to the thermodynamic arrow of time, emerges only at the macroscopic scale when microfacts correlate with macrofacts. It is to be noted that backwards-in-time causal arrows can point only to microfacts and no microfact can do double duty-being both a cause and an effect.
Freebits can thus influence the macroscopic world and cause macrofacts; for instance, a freebit impacting a neuron at a crucial moment might lead to a bad decision being made or avoided.

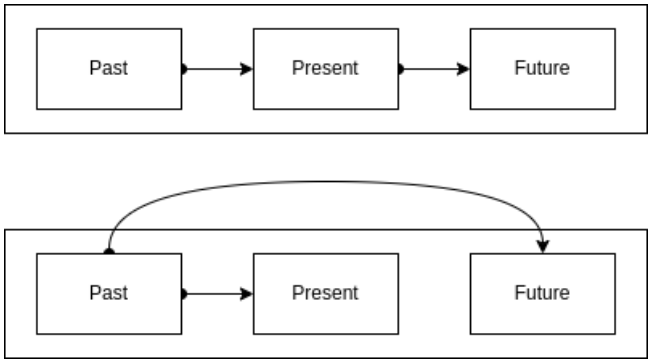

Figure 4. Graphic representation of determinism (above) and the freebit picture (below). According to the freebit picture, it is impossible to estimate the influence of not-yet-measured microfacts on future events when they interact with the world, affecting macrofacts. Image inspired from the one on [22].

7.5 What Are Freebits Even Useful For?

When asked [2] why human choices being randomized by “*hypothetical primordial freebits*” is different from being randomized by the chaotic motion of trillions of molecules inside cells, Aaronson explains [21] that his interest is whether an external observer could build a model of the human through non-invasive scanning that accurately forecasts the probabilities of future choices. If freebits or the molecules inside cells only served as randomization devices, they wouldn’t obstruct such forecasts.
The key idea is that the brain, or other complex systems, might not be sharply divided into a "digital computation part" and a "microscopic noise part," where the former sees the latter purely as a random number source. Thus, cloning a brain and perfectly predicting its behaviour might just be impossible.

8 Can there be consciousness in a digital word?

To conclude this thesis, I would like to add a comment on Chapter 15 of *Reality+: Virtual Worlds and the Problems of Philosophy* [4, p. 270-288]. This was initially intended as a starting point for a discussion with my advisor and colleagues, but I am sure it will serve equally well as a starting point for discussions beyond this thesis. The comment has been slightly revised from its original form to better fit the tone of the thesis, but it otherwise contains all points originally made.

8.1 The three problems

In this chapter, Chalmers talks about **the problem of consciousness**, trying to find a scientific definition and explanation for it. He had previously baptised the problem of explaining consciousness "the hard problem" and since then the expression spread rapidly. Here's a passage of the book that sums it up.

> *When my visual system processes a stimulus in a way that leads me to identify it as red, why do I have a conscious experience of redness? Why is there something it is like to see red?* [4, p. 276]

He then talks about **the problem of other minds**, stating that it is "*hard to know for sure whether silicon machines can be conscious*" because "*it's hard to know for sure whether any entity other than oneself is conscious*" [4, p. 279]. Since it is a personal subjective experience, one can be confident only of their own consciousness. However, this self-assurance doesn't extend to others.

> *How can we know that anyone else has a mind? And how can we know what their minds are like?* [4, p. 279]

Lastly, he discusses whether **machines can be conscious** and does so by introducing the idea of **mind uploading**. He compares three kinds of uploading, with one being effectively cloning (*nondestructive uploading*) and the other two a way of substituting the original brain: *destructive uploading* first clones then destroys the original brain, while *gradual uploading* replaces the brain piece by piece. The consequence of this last type of brain uploading are that the simulated brain would be fully conscious, since how does the consciousness fade away throughout the uploading? Is it slightly reduced at each step of substituting a piece of brain with a piece of simulated brain? Does it just disappear after a certain percentage of neurons have been removed? Is there a key consciousness neuron that turns off the whole conscious experience when removed? Chalmers concludes that "*at least in the special case in which you become a simulated brain by gradual uploading, the simulation will be fully conscious*" [4, p. 286].

8.2 What is consciousness?

When the author asks whether consciousness can be simulated, I believe he fails to provide a satisfactory definition for it. I mean, I know that's the whole issue around consciousness.

For example, he says we can feel and express pain. But can't a simple robot with some sensors that mimic nociceptors [18] and a lightbulb that turns on when those are stimulated also "feel" pain? Do people with congenital analgesia (that cannot feel pain) experience less consciousness than others, just because their body reacts differently to stimuli? Clearly, we would say they experience as much consciousness as anyone else. Then the same could be argued for every other aspect of consciousness, such as speech, memory, logical reasoning, emotions, self awareness and even morality.

The author points out that we tend to explain consciousness by relating it to the physical world, like locating these processes in specific parts of our brain. Sure enough, this is what I just did. But I add that, if we can explain all these elements through physical means, we should be able to simulate them. Yet, simply simulating pain reception, speech, reasoning, etc., does not satisfy me that would be enough to say that we are simulating consciousness as a whole.

8.3 Medium of consciousness

We usually agree that animals are conscious, and that their "amount" of consciousness varies according to their brain size and complexity. An ant seems like an automaton, following pheromones and working for its colony's survival. But other animals exhibit more "human" behaviors, such as emotions, empathy, memory, and self-awareness. Elephants grieve and have burial-like rituals. Crows bring items to "repay" those who feed them. Dolphins recognize themselves in mirrors, showing self-awareness.

Is consciousness merely an issue of medium? In one of our discussions it was argued that humans are conscious because they are natural, while a simulation would be artificial. Imagine we master bioengineering to the point where we can create a brain neuron-by-neuron and a whole nervous system from scratch. Would we deny it consciousness? It would be artificial because it was lab-created, but the components would be the same as ours. If we grant it consciousness, then we could extend this to other artificial simulations. If we deny it, we seem overly protective of our concept of consciousness. Are the circumstances of a brain's creation relevant?

One could argue that a brain is not given an initial learning algorithm while a computer simulation is. However, we accept that the workings of a neural network are largely unknown. Human neurons learn by forming synapses and strengthening connections through repetition, similar to neural networks. If a simulation wasn't given learning instructions, it couldn't utilize its components. Is this difference significant? Is having to tell explicitly to a brain how to function all it takes to differentiate it from a "natural" one?

8.4 Zombies

On the topic of consciousness uploading, the author compares nondestructive, destructive, and gradual uploading. Nondestructive uploading creates a clone of your brain, while destructive and gradual uploading replace the original brain entirely (either all at once or in steps). Gradual uploading is presented as the most consciousness-preserving option. To me, this makes no sense; it's a Theseus' ship situation: by the end of the gradual process, the original brain is gone, leaving only a copy. As mentioned earlier, Chalmers proposes three possible outcomes to a gradual uploading process [4, p. 285-286]:

- **Discrete loss:** you go from full consciousness to no consciousness when a single neuron was replaced.
- **Continuous loss:** at the intermediate point when your consciousness is fading, you're still a conscious being (not a zombie), and you don't display any obvious irrationality. Nevertheless, you're completely out of touch with your own consciousness: you think it's normal when in fact it is fading away.
- **No loss:** your consciousness stays intact at every stage and is present at the end of the process.

To him, the first two outcomes are bizarre and improbable, so he settles on the third and declares that simulated consciousnesses are just as real and valid as "natural" consciousnesses.

Here Chalmers talks about **philosophical zombies**, "a *system that outwardly behaves much like a conscious being, but which inwardly has no conscious experience at all*" [4, p. 271]. A zombie would act like any other human, with the same loquacity, the same perception of pain and the same unique display of character, all while not having any sort of inner experience. Such a zombie can be imagined as a machine that just processes data and is instructed to act as a human, and to believe to be a human. In the continuous loss outcome, he writes that "*you're still a conscious being (not a zombie), and you don't display any obvious irrationality*" while having your consciousness replaced. Is irrationality needed to be a zombie? Does being rational imply being conscious or vice versa? Let's start with

$$\text{Consciousness} \implies \text{Rationality}$$

A conscious being does not have to be rational. To negate this would mean to negate a person consciousness whenever they act irrationally. Is consciousness not continuous over time? Does it have peaks and lows?

Rather we would say that someone's *mental clarity* can change when influenced by different factors (stress, grief, tiredness etc), but we would never say they are not conscious. Thus,

$$\text{Consciousness} \not\Longrightarrow \text{Rationality and Irrationality} \not\Longrightarrow \text{not Consciousness}$$

What about the opposite?

$$\text{Rationality} \Longrightarrow \text{Consciousness}$$

A zombie can be perfectly rational, that is, act logically, without being conscious. Thus once again,

$$\text{Rationality} \not\Longrightarrow \text{Consciousness and not Consciousness} \not\Longrightarrow \text{Irrationality}$$

My conclusion is that, given this explanation, the second option is also probable, at least on the grounds of staying rational throughout the operation.

8.5 Brain uploading and clones

Aaronson also discusses the idea of brain uploading in the F.A.Q. section of "The Ghost in the Quantum Turing Machine", introducing the idea of a perfect predictor.

> **Suppose it were possible to "upload" a human brain to a computer, and thereafter predict the brain unlimited accuracy. Who cares? Why should anyone even worry that that would create a problem for free will or personal identity?** For me, the problem comes from the observation that it seems impossible to give any operational difference between a perfect predictor of your actions, and a second copy or instantiation of yourself. If there are two entities, both of which respond to every situation exactly as "you" would, then by what right can we declare that only one such entity is the "real" you, and the other is just a predictor, simulation, or model? [1, p. 16]

Would a copy act like its original? What does it mean to act like it? If your brain was cloned and placed in the world, clearly it would act differently: to act exactly the same it should have your same physical extension, that is to be in the exact same place you are at the exact same time, and experience the same physical phenomena at the same time. In the real world that is impossible: would your clone follow you and stand a little bit to the side? Would its movements be simply delayed? What if a fly goes into your eye but not your clone's? Different experiences lead to different behaviors after all, so your life and your clone's life would diverge completely from that moment of creation.

Since such an experiment would be impossible in the physical world, we could just simulate two copies of the same nervous system and subject them to identical artificial stimuli [1]. Even then, it appears that the brain (and more generally, neurons) doesn't necessarily process the same stimuli identically each time [23, 28]. So it is not just different experiences that define one or the other consciousness, but it is the brain itself, processing inputs differently each time, that makes each consciousness unique. After creating a clone of your brain, your lives will diverge completely. Multiple simulated consciousnesses of the same person could simulate the different decision outcomes.

In conclusion, one cloned consciousness is not less the original than the original itself; a consciousness can express itself in multiple ways but is limited to one expression at a time. A cloned consciousness will certainly not act like the original, but represents a different expression of its potential. The cloned consciousness is the same as the original because it is different from the original. A different expression of the original.
In other words, you are you because you are never you [2].
Is consciousness just freedom from necessity?

[1] While the input and the analysis of the outputs should be done via a machine, the simulations should be run on "meaty" brain hardware.

[2] More appropriately, you are you because **you are not any of the other alternative possibilities of yourself**.

9 Bibliography

[1] S. Aaronson. The ghost in the quantum turing machine. 2013.

[2] anonymous on LessWrong. Quotes and Notes on Scott Aaronson's "The Ghost in the Quantum Turing Machine" — LessWrong. https://www.lesswrong.com/posts/NnzYPN7mHDQ7hpeTc/quotes-and-notes-on-scott-aaronson-s-the-ghost-in-the?commentId=uE3JXTDMMak9AEHPY.

[3] H. Bergson. *Time and Free Will*. Humanities Press, New York, 1910.

[4] D. J. Chalmers. *Reality+: Virtual Worlds and the Problems of Philosophy*. W. W. Norton, New York, 2022.

[5] W. contributors. Free will — wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Free_will&oldid=1221443044, 2024. Accessed: 2024-06-15.

[6] B. Doyle. Determinism. Accessed: 2024-06-15.

[7] B. Doyle. Libet experiments. Accessed: 2024-06-20.

[8] B. Doyle. The standard argument against free will. Accessed: 2024-06-15.

[9] P. Forrest. The Identity of Indiscernibles. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition, 2020.

[10] T. Hoffmann and C. Michon. Aquinas on free will and intellectual determinism. *Philosophers Imprint*, 17, 05 2017.

[11] D. R. Hofstadter. *I Am a Strange Loop*. Basic Books, New York, NY, USA, 2007.

[12] P. V. Inwagen. *An Essay on Free Will*. Oxford University Press, New York, 1983.

[13] Julia Jorati. Gottfried Leibniz: Philosophy of Mind | Internet Encyclopedia of Philosophy. Accessed: 2024-06-15.

[14] S. M. Kaye. Why the liberty of indifference is worth wanting: Buridan's ass, friendship, and peter john olivi. *History of Philosophy Quarterly*, 21(1):21–42, 2004.

[15] G. W. Leibniz and S. Clarke. *Exchange of Papers Between Leibniz and Clarke*. Jonathan Bennett, 1715.

[16] B. W. Libet. Can conscious experience affect brain activity? *Journal of Consciousness Studies*, 10(12):24–28, 2003.

[17] W. Penfield and C. Symonds. *Mystery of the Mind: A Critical Study of Consciousness and the Human Brain*. Princeton University Press, 1975.

[18] D. Purves, G. J. Augustine, D. Fitzpatrick, et al., editors. *Neuroscience*. Sinauer Associates, Sunderland (MA), 2nd edition, 2001. Available from: https://www.ncbi.nlm.nih.gov/books/NBK10965/.

[19] N. Rescher. Choice without preference. a study of the history and of the logic of the problem of "buridan's ass". *Kant-Studien*, 51(1-4):142–175, 1960.

[20] A. Schurger, J. D. Sitt, and S. Dehaene. An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences*, 109(42):E2904–E2913, 2012.

[21] ScottAaronson on LessWrong. Quotes and Notes on Scott Aaronson's "The Ghost in the Quantum Turing Machine" — LessWrong. https://www.lesswrong.com/posts/NnzYPN7mHDQ7hpeTc/quotes-and-notes-on-scott-aaronson-s-the-ghost-in-the?commentId=CqWGbXEqjaMrc6r3L.

[22] shminux on LessWrong. Quotes and Notes on Scott Aaronson's "The Ghost in the Quantum Turing Machine" — LessWrong. https://www.lesswrong.com/posts/NnzYPN7mHDQ7hpeTc/quotes-and-notes-on-scott-aaronson-s-the-ghost-in-the.

[23] T. Stephani, G. Waterstraat, S. Haufe, G. Curio, A. Villringer, and V. V. Nikulin. Temporal signatures of criticality in human cortical excitability as probed by early somatosensory responses. *Journal of Neuroscience*, 40(34):6572–6583, 2020.

[24] Stephen P. Thornton. Freud, Sigmund | Internet Encyclopedia of Philosophy. Accessed: 2024-06-15.

[25] The Editors of Encyclopaedia Britannica. Identity of indiscernibles, 2024. Accessed: 2024-06-01.

[26] The Editors of Encyclopaedia Britannica. Jean Buridan, 2024. Accessed: 2024-06-02.

[27] E. Vailati. *Leibniz & Clarke: A Study of Their Correspondence*. Oxford University Press, 1997.

[28] W. A. Wybo, B. Torben-Nielsen, T. Nevian, and M.-O. Gewaltig. Electrical compartmentalization in neurons. *Cell Reports*, 26(7):1759–1773.e7, 2019. https://actu.epfl.ch/news/the-way-a-single-neuron-processes-information-is-n/.