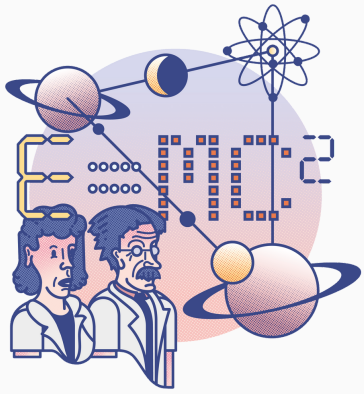

The Neuroscientist Project

Younjin Kim

SUPERVISED BY HAZEL (OXFORD PHILOSOPHY & THEOLOGY GRADUATE AND MASTER OF PHILOSOPHY IN ANTHROPOLOGY)

The Neuroscientist: 4 thought experiments on moral responsibility and agency

Written by Jin Kim, Finalised on September 14th 2022

I. Introduction

This paper hopes to investigate the nature of free will and moral responsibility using a series of thought experiments which closely resembles consensus on determinism today. Rather than trying to collate theories of free will or present the wide variety of philosophical debates in history, this paper presents a series of iterations on one key thought experiment, *The Neuroscientist*, in order to bring together different disciplines within the study of philosophy. *The Neuroscientist* attempts look at an old issue in a new light – instead of evaluating the validity of free will or its compatibility with determinism, I will be examining the importance and integration of the illusion of free will in society.

II. Abstract

The Neuroscientist is closely adapted to, and inspired by, Harry Frankfurt's *Black and Jones*[1] analogy, which is a counterexample to highlighting the importance of the consideration of the Principle of Alternate Possibilities (PAP)- the notion that an agent has free will if he 'could have done otherwise' inside the free will debate. Frankfurt's counterexamples have been breaking in the free will debate by stimulating discussion around the true significance of PAP to the notion of freedom and its respective importance to moral responsibility. *The Neuroscientist* attempts to add to what Frankfurt started to discussed with his examples by discussing the nuances in causation around an agent in relation to free will.

Frankfurt's analogy involves Black, a neuroscientist, who wants Jones to perform a certain action, but prefers to avoid unnecessary intervention. Black waits until Jones is makes his decision and therefore does nothing until it is evident that Jones is not going to do what Black wishes. As it turns outs, Black never has to intervene as Jones, for his own reasons, performs the actions which Black wants him to perform.[2]

The power of *The Neuroscientist* as a thought experiment comes from the fact that it presents a model very similar to that of the world we believe we live in – where we perceive ourselves as free beings, which may be a gross illusion. Using this third person, 'omniscient narrator' perspective, it is easier to gage what man values in both moral responsibility and a sense of freedom.

[1] See Frankfurt, Harry G. (1983). *What We Are Morally Responsible For.* Published by Hackett Publishing Company.
[2] Sartorio, Caroline. (2016). *Frankfurt-Style Examples.* Published by Routledge. Retrieved from https://sartorio.arizona.edu/files/Routledge.pdf

III. What is *The Neuroscientist?*

Throughout their entire life, an agent, A, has had their actions, thoughts and brain activity monitored constantly by a neuroscientist, N. N can do one of two things: a) leave A to do as she wishes, or b) alter the neurology of her brain to change his thoughts and actions. Originally, A has no knowledge of N, and believes she is completely morally autonomous.

Each iteration of *The Neuroscientist* will attempt to come to the conclusion of two questions:

1. How does A, N and their relationship with each other and the outside world affect our perception of free will?
2. What does this show about societal standards of moral responsibility?

IV. Iterations

Iteration A – Intervention

Scenario 1 (S1): A goes to carry out action X. Since N desired A to carry out X, he does not need to alter the neurology of A. A carries out X.

Scenario 2 (S2): A goes to carry out action Y, however N desires A to carry out action X, so therefore alters the neurology of A so that A carries out X.

The two scenarios are clearly different due to the intervention of N in A's desire to carry out an action, however they have more similarities than first thought after closer examination. Both Scenario 1 and 2 establish a situation where there is no 'ability to do otherwise', and therefore the Principle of Alternate Possibilities (PAP) should mean that A does not have moral autonomy in either scenario. It seems rational to believe that A is 'more free' in S1 than S2, due to the lack of N's intervention is A's neurology, but this belief is in fact, irrational primarily for two reasons. The first argues that A is not free in both scenarios, whilst the second advocates the view that A is free in both these cases.

The first is that, in S1, the fact that N desired X and A also happened to desire X comes down to chance. Gettier makes the point that there is an aspect of being 'accidently correct' such as guessing the date of the Summer Solstice or hedging a bet on when the Battle of Hastings took place, as well as the core principles of knowledge, having a justified, true belief.[3] As with knowledge, the fact that A and N just so happened to desire the same action X creates this

[3] Gettier proposes that there is more to knowledge than the traditional idea of having a justified, true belief. He uses an example in which a man, Smith, has strong evidence that Jones has a Ford, and also, is totally ignorant on the whereabouts of Brown. Smith thus decides to make a proposition that Either Jones has a Ford, or Brown is in Boston (ie he chooses a random location). Of course, Smith is completely justified in believing this proposition. However, it happens that Jones does not has a Ford and was in fact renting one, and Brown so happens to be in Boston, meaning this statement is true, but can be perceived in a different way to initially thought, thus creating the concept of 'accidental knowledge.'

For reference, see Gettier, Edmund. (1963) *Is Justified True Belief Knowledge?* Published by the University of Massachussetts, Amherst.

sense of 'accidental free will', where there is a sense that there is also a high probability that either could have chosen differently, undermining the authenticity of freedom in S1. Perceived PAP in this iteration can circulate around both A or N. It comes from the idea that A had the potential to carry out a different action, which so happened to be desired by N, or that A could still have carried out the same action but N could have instead desired this action (and therefore no intervention was needed).

The second reason that A is not 'more free' in S1 than S2 is down to the awareness of A in both scenarios. Although in both cases PAP fails and there is no ability for A to do otherwise, it can be argued that A is a free agent in both scenarios due to Kane's plurality of conditions. In both S1 and S2, A is confronted with two options: either to perform action X, or not to perform action X, and in both scenarios A believes that they have the ability to voluntarily choose X or any other action. This belief means that in terms of free will, A believes she is free in both cases, and therefore has culpability for her action.

Iteration A of *The Neuroscientist* shows that there is more to moral responsibility and free will than just the truth of determinism and our neurological behaviour. Both arguments against the distinguishing of S1 and S2, though coming to different conclusions, highlight that the final outcome of each is the same; that A performs X with the perception that she is free. Much like A in this iteration, society accepts free will at face value without certainty that this free will, in fact, exists. The idea that moral autonomy should be preserved for the sake of responsibility and societal functioning is already embedded into society. As Smilansky suggests with the idea of illusionism[4], we as agents tend to believe in our own freedom, and perhaps this view should not change. Iteration A shows us that our perception of free will mainly comes from an internal justification of such a concept – we *feel* able to do as we wish, even if this is not the case. This iteration establishes that there is a relationship between moral responsibility and internal justification which may have been ignored previously in the free will debate. The idea that free will is an illusion is not only reinforced on a societal level, but also perhaps internally within an agent. So, instead of Smilansky's illusionism being brutish and cruel, it alters perception as this illusionism is already happening on an individual level.

Iteration B: The Primary Agent's Knowledge

A gains awareness of the fact that N has the potential to, and history of, changing the neurology of A's brain in order to dictate her actions.

Developing on the established importance of awareness in the idea of moral responsibility realised from Iteration A, this alteration to *The Neuroscientist* more closely examines the impact of knowledge on A on an individual level.

As previously established, awareness of N impacts moral responsibility, but not the necessarily A's freedom as an agent. Since there is nothing A can do to prevent or detect when

[4] Kane, Robert. (2005). *The Oxford Handbook of Free Will.* Published by Oxford University Press.

N is changing her neurology, the essence of A's freedom does not fundamentally change with gained realisation of N's abilities. Iteration A establishes that in the event of the same action being performed with and without intervention, A can only be free in both cases due to the lack of awareness, it is rational to deduce that in Iteration B, A does not possess free will.

Despite the essence of A's freedom being fundamentally consistent with the first iteration, the authenticity of it has changed significantly. Gained awareness of N changes A's internal perception of free will on a psychological level. Since A cannot prevent nor detect N, it remains true that sometimes A's actions is being determined, and sometimes they are not. However, A's gained awareness of N may change the frequency of which A is being determined by N, causing her to act in the way in which she believes N desires. The fact that A has the potential to create a mental persona of N means that she might modify her actions in a way which either N desires, or means more intervention is needed.

Iteration B1: What if as a result of A's gained awareness, N never has to intervene again?

B1 would mean that A would still believe she is being determined by N, though this is not the case – has the authenticity of A's freedom changed? Since A fundamentally still has awareness of N, the idea that she can introspectively blame or praise N whenever she wishes has been established. Perhaps B1 explores a more psychological and ethical approach to the subject of free will and its relation to knowledge; A could choose to blame N for her immoral actions, or believe N is controlling her actions which are irrational. Since A is now not determined by N, perhaps B1 elucidates more on the psyche of A rather than the nature of her freedom itself.

In terms of moral responsibility, the nature of A's new knowledge that makes the impacts of this iteration difficult to quantify – the ambiguity of when N is affecting A. This ambiguity complicates the situation in terms of free will, but also perfectly depicts the controversy around the free will debate today, since uncertainty is present in both cases. As established by B1, A has the ability to introspectively praise or blame N for a given action. However, this adds a modicum of doubt that A is not truly acting to her desires, which hinders her perceived moral responsibility. Ultimately, since A cannot pinpoint exactly when N is intervening with her actions, this feeling of chance makes A feel less culpable on an individual level.

Iteration B, along with the findings of many psychological experiments such as that of Vohs and Schooler (2002)[5] demonstrates that the less an agent believes in their moral autonomy, the less the feel an obligation to act morally and the breakdown of moral obligation. Determinism which has been rising in science throughout the modern era as well as the neuroscientist are the same in this scenario, making people doubt their moral responsibility, highlighting the threat of determinism. Instead of this idea that free will is the illusion, Iteration B suggests that perhaps determinism is the illusion; a façade which one uses to avoid culpability for their actions.

Iteration C: Indeterminism and Chance

[5] Cave, Stephen. (2016). *There's no such thing as free will.* Published by The Atlantic. Retrieved from https://www.theatlantic.com/magazine/archive/2016/06/theres-no-such-thing-as-free-will/480750/

N uses the flip of a coin to decide when he decides to interfere A's neurology.

N making his decisions due to luck is reminiscent of modern day findings in quantum physics and the fact that indeterminism is embedded into science on a microscopic levels. Though this iteration is not to be used as an explanation for quantum physics, a background knowledge is essential to understanding the indeterminate nature of our world. Heisenberg's uncertainty principle means that it is impossible to know both the position and speed of a particle with perfect accuracy, meaning a guess must be made about one of these two variables. Furthermore, it has been proven that particles do unusual or irregular things when not being watched by the human eye, and that regularity or predictable events are performed when particles are under direct inspection. This sub-branch of physics opens the door to suggestions that the natural world is not, in fact, determined, much like N flipping a coin to decide whether to alter A's initial decision.

Though quantum physics has proved indeterminism on a subatomic level, there are still huge implications for the free will debate. The fact that A's fate is being decided by a flip of a coin rather than N's rational decision affects the way we perceive A's freedom. Rather than seemingly deterministic or control, the element of chance makes N's intervention seem less deliberate. It builds a perception that although A's neurology is still being altered, A is not being controlled as seen as in the other alterations.

The probability aspect of this iteration also adds to the relationship between A and N, though unknown to both of them. A balance of power is restored since the N's free will has been replaced with a flip of a coin. Power and control have been removed from both of them, arguing altering the perception of A's freedom yet again. Since N's intervention is not a deliberate, rational action, A's inevitable fate feels more natural. The lack of rationality behind the (in)determinism of N as well as the world around us presents this idea that free will can involve causation, but not the determinism by a rational agent – just how laws of science and The Matrix are not the same deterministic causes.

How does indeterminism affect moral responsibility, when, it still arguably leads to the fact we have no moral autonomy? Kane argues that indeterminism does not equate to a lack of freedom, but rather they are inseparable from each other. Indeterminacy within in a decision does not hinder a sense of control, but rather is part of the process and in itself a cause. Kane addresses the fact that indeterminism and causation are not mutually exclusive – in fact, the probability aspect of N's decision arguably makes A more free than in any of the other iterations. The lack of rationality behind N's decision whether or not to manipulate A, makes A more morally responsible, as there is less culpability and reasoning on N's front. Iteration C makes the point that the illusion of free will is bolstered, not hindered by the deterministic and indeterminstic laws of science, since the source is seen to be natural and not rational.

Iteration D: Thoughts vs Actions

This iteration does not involve a change in the original thought experiment, but examines the model from a different perspective: it is repeatedly stated that N controls A's neurology,

but what impact does this have on her actions? The fact that N does not have explicit control over A's actions paves the way for the will to bridge the gap between thought and action. This unique form of causation from thought to action is perhaps what free will libertarians describe as agent-causation.[6] Libertarians, who believe that free will is incompatible with determinism, but still exists, often use agent-causal theories to prove the existence of free will. Agent-causation is the idea that there is a process of causation within an agent which is irreducible, but significant nonetheless. The agent herself acts as an additional factor or cause which explains why some actions happen and others don't, regardless of chance.

Agent-causation theorists could argue that N having the potential to alter A's neurology does not limit or hinder her free will in any way, since the will is within the agent herself. What agent-causation is trying to push forward about moral responsibility is how it is part and parcel of being a person, and that this sense of internal causation is what binds free will to our responsibility as moral beings. The threat of determinism to the illusion of freedom, in this case, is neutralised – the deterministic nature of neurology is significant to the agent's freedom due to the ontological gap between thought and action.

V. Conclusion

The four iterations above have been different responses to the free will debate, but have all discussed the idea of free will as an illusion in the modern era. Illusionism does not eliminate scope for freedom in the slightest, but rather, shows that this illusion is an interpretation of human rationality within the natural world. Deterministic and indeterministic laws are not threats to the free will illusion, but rather further stimulate ideas on what autonomy truly is and why it is more than just causation.

The Neuroscientist has opened the potential to investigate the relationship between factors which are interdisciplinary, but are important to the free will debate nonetheless. From the psychological impact of awareness on free will to the quantum physics of indeterminism, I hope this thought experiment has widen scope for investigate the nature of these ideas and their relation to each other. The first two iterations examine the perspective of A herself, and how the unclear authenticity of our freedom adds to this idea that free will as an illusion requires ambiguity to thrive. Iterations C and D then build on this ambiguity by examining the impact of chance as well as the unclear relation between thought and action to consolidate the idea that physics does not impede the idea of free will, but further supports the idea that free will is in fact a social concept rather than a natural one.

The free will debate is more than just the compatibility of determinism and freedom; *The Neuroscientist* emphasises that nuances in causation, perception of psychology and knowledge all play into the synthetic idea of free will which man has created.

---

[6] See Kane, Robert. (2005). *The Oxford Handbook of Free Will.* Published by Oxford University Press.

VI. Bibliography

1. California Institute of Technology. (n.d). *What is the Uncertainty Principle and Why Is It Important?* Published by Caltech Science Exchange. Retrieved from https://scienceexchange.caltech.edu/topics/quantum-science-explained/uncertainty-principle
2. Cave, Stephen. (2016). *There's no such thing as free will.* Published by The Atlantic. Retrieved from https://www.theatlantic.com/magazine/archive/2016/06/theres-no-such-thing-as-free-will/480750/
3. Clarke, Randolph, Capes, Justin & Swenson, Philip. (2021 Edition) *Incompatibilist (Nondeterministic) Theories of Free Will.* Published by The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/incompatibilism-theories/#3
4. Edmonds, David & Warburton, Nigel. (2012). *Philosophy Bites Again.* Transcript published by Oxford University Press. Retrieved from https://s3.us-west-2.amazonaws.com/secure.notion-static.com/e92790fc-f8b6-4e6e-9516-866d9668d074/bitesagain.pdf?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Content-Sha256=UNSIGNED-PAYLOAD&X-Amz-Credential=AKIAT73L2G45EIPT3X45%2F20220822%2Fus-west-2%2Fs3%2Faws4_request&X-Amz-Date=20220822T132046Z&X-Amz-Expires=86400&X-Amz-Signature=178f0e8712d665197bd5010a162bde39b8ef37936e54eccfbecd496dd0509ce2&X-Amz-SignedHeaders=host&response-content-disposition=filename
5. Finlay, Stephen and Schroeder, Mark. (2017 edition). *Reasons for Action: Internal vs. External.* Published by The Stanford Encyclopedia of Philosophy. Retrieved from https://plato.stanford.edu/entries/reasons-internal-external/
6. Frankfurt, Harry G. (1983). *What We Are Morally Responsible For.* Published by Hackett Publishing Company.
7. Kane, Robert. (1996). *Freedom, Responsibility and Will-Setting.* Sourced by Philosophical Topics, Vol. 24, No. 2. Published by University of Arkansas Press. Retrieved from https://www.jstor.org/stable/43154237
8. Kane, Robert. (2005). *The Oxford Handbook of Free Will.* Published by Oxford University Press.
9. Kane, Robert. (n.d). Reflections on Free Will, Determinism and Indeterminism. Published by The Determinism and Freedom Philosophy Website. Retrieved from https://www.ucl.ac.uk/~uctytho/dfwVariousKane.html
10. Lemos, John. (2011). *Wanting, Willing, Trying and Kane's Theory of Free Will.* Sourced from Dialectica, Vol. 65, No. 1. Published by Wiley. Retrieved from https://www.jstor.org/stable/42971235
11. Sartorio, Caroline. (2016). *Frankfurt-Style Examples.* Published by Routledge. Retrieved from https://sartorio.arizona.edu/files/Routledge.pdf
12. Smith, Quentin. (1994). *Time, Change and Freedom.* Published by Routledge.
13. Talsma, Tina. (2012). *Free Will and Divine Omniscience.* Published by Florida State University Libraries. Retrieved from https://s3.us-west-2.amazonaws.com/secure.notion-static.com/8f029216-e46b-4f8c-9974-c493614d38fe/view.pdf?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Content-Sha256=UNSIGNED-PAYLOAD&X-Amz-Credential=AKIAT73L2G45EIPT3X45%2F20220822%2Fus-west-2%2Fs3%2Faws4_request&X-Amz-Date=20220822T132616Z&X-Amz-Expires=86400&X-Amz-

Signature=36e79617a75a0ffc62035a7206238a23445a8f719a67682fcf701e00b7bbca2c&X-Amz-SignedHeaders=host&response-content-disposition=filename
14. Van Inwagen, Peter. (2014). *Metaphysics.* Published by Routledge.