# Reimagining Free Will and Moral Responsibility

Nietzsche's Practical Approach

Yifan Li

A thesis submitted for the degree of Doctor of Philosophy

Department of Philosophy

University of Essex

October 2025

# Abstract

Most scholars engage in the free will and moral responsibility debate by examining its relationship with determinism. However, this debate remains trapped in conceptual analysis, showing no sign of resolution. Inspired by Strawson's *Freedom and Resentment* (1962), I argue that Nietzsche offers a method to move beyond the compatibilism/incompatibilism dualism by analysing the development of conscience in response to moral expectations and punishment. In this account, conscience and guilt are corresponding attitudes to moral expectations and to be a free agent is a matter of competence, which hinges on one's ability to assume responsibility for oneself. I first focus on moral responsibility, suggesting that Nietzsche sees morality as a natural phenomenon evolving alongside moral practices. Through an analysis of its evolution, I argue that Nietzsche does not reject moral responsibility but instead advocates for a version promoting health, rooted in individuals' strong will to create values and hold themselves responsible based on pride and confidence in their ability to rectify misdeeds. To develop a positive account of Nietzschean free will, I claim that he explicitly defines free will as "the will to self-responsibility" (*TI*, Skirmishes, 39), a form of strong will. The elements shaping conscience and value creation—autonomy and *amor fati*—constitute Nietzsche's positive interpretation of free will. To support this, I analyse the origins of "freedom" and "the will" and examine how Nietzsche builds on Stoic ideas of fate and free will by emphasising the active embracing of suffering as essential to autonomy. Nietzsche's practical understanding of free will as a strong will in the practice of *amor fati*—one that actively creates its own values amid the constraints of existence—challenges traditional notions of free will and offers a new approach to the determinism-free will/moral responsibility debate. This also has the potential to free contemporary debates from endless conceptual analysis.

# Acknowledgements

Four years have passed in the blink of an eye. Now it seems like a mere instant, yet the weight of that instant cannot be captured by the simple measure of "four years".

My encounter with Nietzsche was purely accidental. Had I not taken the Nietzsche module during my master's studies, my impression of him might have remained forever on the level of poetic rhetoric—never reaching the thinker who stands beyond the poet's temperament. My encounter between Nietzsche and Strawson, however, was even more unforeseen. I still remember that afternoon in the Mingfu Library in Shanghai: as I read how Strawson, starting from our everyday emotions and responses, drew moral responsibility down from the heights of abstract metaphysics into the immediacy of lived reality, Nietzsche's descriptions of punishment and responsibility in *On the Genealogy of Morals* suddenly flared to life. Yes—resentment! Strawson wrote "Freedom and Resentment"; Nietzsche traced how *ressentiment* in Christian morality evolved into a sense of guilt. Their aims were different, their eras far apart, and yet there must be a hidden thread connecting them—one that reveals, through layer upon layer of interpretation, the fundamental meaning of moral responsibility in our everyday lives.

For four years I have been trying to render this thread more clearly. Owing to my own limitations, what I have achieved can only reach so far. But even now, when I hold in my hands all that I have written, I remain struck by how Nietzsche's understanding of free will and moral responsibility breaks through the confines of tradition. Perhaps this thought can never be scientifically verified, but at least Strawson lends to Nietzsche's fervent rhetoric a finer texture of analysis and annotation. As Nietzsche once said, humans would rather seek meaninglessness than submit to meaninglessness at all. For us, even if our inquiry into moral responsibility and free will may forever remain an inquiry, are we not still blessed—so long as we can still grasp our joys and sorrows, our love and resentment?

As I said before, many coincidences have brought this thesis into being. There are always those one must thank—their names rest like a piece of sugar beneath the tongue, too sweet to spit out, yet impossible not to utter. And so I must write them down. Just as I have learned to look calmly upon the great lies of the world, the other side of that coin is this: the gratitude I owe to each of these names. Without them, how solitary this journey would have been.

My deepest thanks go to Fiona, my supervisor, who first recognised the potential of my proposal and guided me onto the proper path of doctoral research. It is hard to imagine how much effort it must have taken to supervise a non-native English speaker through so many meetings and countless rounds of revision. If ever given the chance, I hope I can influence others as she has influenced me—with seriousness and warmth that never wane.

I also thank Irene, my interim supervisor, who took on a task beyond her original duty. Though our research fields differ greatly, and our modes of thinking diverge, she still guided me meticulously through the final stages of revision. Without her, many of my thoughts might have drifted away—never settling into the weight of words that could be seen and understood by others.

My gratitude extends to Beatrice, Chair of our department, who created such a nurturing atmosphere for my study at Essex. As my internal examiner, her questions during my viva left me speechless—but they also became the turning point that transformed my thesis from something crude into something truly mine.

I thank Paul, my external examiner. I never imagined that the author whose work had so deeply moved me during my master's studies would one day read my own. He inspired me to approach Nietzsche from a psychological perspective—without that inspiration, I might still be wandering in the circle of metaphysics.

Thanks also to Steven, a long-time member of my panel. Though not my supervisor, he read and commented on my chapters as if I were his own student, helping me to think beyond the surface of Nietzsche and Strawson.

To Wayne, whose insight into Frede helped me trace the Stoic roots of the concept of free will—thus grounding the idea of the will to power as a will to freedom; to Daniel, with whom I mused on the possibility of a freedom without moral responsibility; and to Henri, who introduced me to Bernard Williams's reflections on shame and morality in ancient Greece—thank you. And to all the members of my committee, for reminding me that philosophy is never a solitary pursuit.

My heartfelt thanks go also to my parents. It is their steadfast support that has allowed me to see a world I could never have seen on my own. I possess no particular talent, nor am I especially diligent; while my peers have already made their mark in their respective fields, I am still sitting at a quiet desk, reading and writing things that seem untimely. For this I am profoundly grateful, though I have no way to repay it. I can only keep walking, step by step, hoping their faith in me will not vanish into nothingness.

Lastly, my deepest gratitude to my partner, Yuwan, who has walked beside me all this way. She has witnessed everything—from my proposal revisions, to the PhD application, to the writing and final edits of this thesis. Zen philosophy says: "When one drinks water, one knows for oneself whether it is warm or cold." In Western philosophy, the "problem of other minds" asks whether we can ever know that another has a mind. But I am not troubled by such doubts. She understands me—and she strives to understand what I, too, am trying to understand. This is no behaviourist cliché; even without words or gestures, she has always been here.

There is still much left to be said, but I shall end here. I only hope that the words I have written will never betray the thoughts that gave them life—now, as in the future.

# Contents

# Introduction

Nietzsche is probably best known for his self-proclaimed "immoralism", a concept subject to diverse interpretations. One possible interpretation is scepticism or even nihilism, which entails questioning or outright rejecting any coherent moral framework. Another interpretation, though acknowledging the existence of morality, considers morality per se to be fundamentally detrimental to humanity, suggesting that we ought to adopt an "immoralist" stance and contribute to the denial of moral values. Consequently, throughout the 20th century, numerous philosophers—such as Arthur Danto, Brian Leiter, and Walter Kaufmann—questioned whether Nietzsche presents a coherent and affirmative ethical framework. Indeed, while Nietzsche frequently asserts that his aim is the destruction of ideals, he simultaneously advocates for certain values and characteristics he deems preferable. He repeatedly proclaims certain ends and goals worth achieving. Among these, the concepts of moral responsibility and free will are particularly important. This is because Nietzsche's revaluation of morality—or his critique of moral values, which I take to be the true essence of his "immoralism"—requires, as he states, "a knowledge of the conditions and circumstances out of which they have grown, under which they have developed and shifted" (*GM* P: 6). The cultivation of morally responsible agency in humans, alongside the conventional conception of free will often invoked to justify it, is therefore indispensable to this process. As Nietzsche himself emphasises in his praise of the sovereign individual, described as "the ripest fruit" (*GM* II: 2) of the development of morality and endowed with the capacity for free will, we may still rely on these two concepts as practical constructs, rather than seeking to dissolve them entirely—even if they are far more intricate than traditionally depicted. Whatever the true nature of moral responsibility and free will, Nietzsche remains committed to certain defining aspects of human agency that he deems essential. Therefore, this dissertation will focus on examining the specific notions underpinning Nietzsche's unique commitments, namely moral responsibility and free will.

Debates regarding Nietzsche's views on free will and moral responsibility have persisted for decades. Much of this controversy originates from broader disputes within the traditional discourse on free will and moral responsibility. The traditional debate on free will is fundamentally centred on the dilemma of determinism. A central premise of the traditional debate is that, with few exceptions, the "modern tradition" in philosophy has equated "moral agency" with "free agency," as free will has traditionally been regarded as a prerequisite for moral responsibility. On one hand, free will appears incompatible with determinism, as the latter implies that every event or action is the inevitable consequence of prior causes, thereby intuitively conflicting with our understanding of free will. Conversely, free will also appears incompatible with indeterminism. If an action is not determined by other factors, its occurrence appears to be merely a chance event, which cannot be attributed to any agent. Therefore, determinism presents us with a dilemma. Following this, free will has been conceived as a kind of power to control one's choices and actions. In other words, free will requires one's choice and action being up to her in the sense that she could have done otherwise or she is the ultimate source of her action. Robert Kane (2005) summarises two primary understandings of the notions of free will in contemporary literature. The first states that for an individual to possess free will, she must have the capacity to act otherwise than she actually does. This approach is referred to as "alternative possibilities" sense of free will (AP free will). The second does not centre on whether the future remains open. For an agent to be free, the origin of their choices must be located within themselves, thereby establishing them as the ultimate source of their actions. This conception of free will is referred to as "sourcehood conception" of free will. Compatibilism is the view that free will can be compatible with determinism. Some compatibilists argue that causal determinism does not negate our ability to do otherwise. Others reject the necessity of alternative possibilities for meaningful free will. What is required is to be the ultimate source of one's actions. In contrast, incompatibilism asserts that free will and moral responsibility are fundamentally incompatible with causal determinism. Some incompatibilists, known as libertarians, hold the view that we possess free will, which implies that the world is not wholly deterministic. Others, known as hard determinists, reject the existence of free will and moral responsibility, asserting that human actions are entirely governed by external factors. Both groups of incompatibilists cannot affirm that we have free will and that the world is deterministic at one time. They argue that determinism is a threat to the necessary condition for free will. On one hand, determinism rules out the existence of alternative possibilities. This argument can be formalised as follows:

*Alternative Possibilities Incompatibilist Argument*

1. Free will requires alternative possibilities.
2. If determinism is true, no one has alternative possibilities and could have done otherwise.
3. Therefore, free will is incompatible with determinism.

On the other hand, if determinism is true, then past events, combined with the laws of nature, would be the ultimate source of every human action. Therefore, any given action has its source in causal antecedents outside of the agent. This reasoning is captured in the following argument:

*Sourcehood Incompatibilist Argument*

1. Free will requires one to be the ultimate source of one's action.
2. If determinism is true, no one can be the ultimate source of her action.
3. Therefore, free will is incompatible with determinism.

Broadly speaking, the traditional debate considers the attainment of free will and moral responsibility to depend on fulfilling certain preconditions, like having alternative possibilities or being the ultimate source of one's action. Once these conditions are met, the challenge posed by determinism can be circumvented, enabling the agent to possess free will and thereby qualify as a moral agent. As previously noted, most Nietzsche scholars have examined the debate on free will and moral responsibility from the traditional perspective, trying to identify these preconditions and their relationship with determinism. Brian Leiter is among the most prominent proponents of this approach. In *Moral Psychology with Nietzsche* (2019), Leiter summarises his earlier arguments, asserting that Nietzsche cannot hold any kind of responsibility and free will. His argument is based on Nietzsche's distinctive form of fatalism, his scepticism regarding the causal efficacy of the will, and his epiphenomenalism concerning conscious mental states related to actions. All of these are shown to undermine both compatibilist and incompatibilist views of moral responsibility and free will. Leiter's interpretation of Nietzsche's philosophy is grounded in what he calls "Doctrine of Types" (Leiter, 2002: 91). According to the "Doctrine of Types," "each person has a fixed-psycho-physical constitution, which defines him as a particular type of person" (ibid.). He calls these constitutions "natural" facts or "type-facts". They are "facts about the person's unconscious drives or affects‖" (ibid.), which determine an individual's life trajectory.

Leiter offers two different models of willing: one is the "epiphenomenal" (Leiter, 2007: 13) model and the other is the "secondary cause" model (ibid.). According to the epiphenomenal model, conscious mental states play no role in the causal chain bringing about actions; rather, "conscious states are simply effects of underlying type-facts about [a] person, and play no causal role whatsoever" (Leiter, 2002: 92). In contrast, the secondary cause model acknowledges a role for conscious mental states in the causal chain leading to action, yet denies thay they are the ultimate cause. Type-facts cause "the experience of willing and then the will causes the action" (Leiter, 2007: 13). Leiter supports his models with three arguments: (1) the phenomenology of willing, (2) confusing cause and effect, and (3) the real genesis of action (Leiter, 2019). According to Leiter, Nietzsche's texts are ambiguous as to which one of the views about the epiphenomenalism he embraces. Therefore, he turns to Libet (1985) and Wegner's (2002) empirical evidence to support his claim that the epiphenomenal model is the primary claim Nietzsche defends. If Libet and Wegner are correct, this provides strong justification for interpreting Nietzsche as holding the epiphenomenal model which is "(a) supported by his texts and (b) most likely to be correct as a matter of empirical science" (Leiter, 2007: 13).

Several scholars attempt to circumvent the metaphysical issue while still trying to identify a precondition for moral agency—one that could be compatible with determinism and thereby permit free will. For example, Aaron Ridley (2007) convincingly argues that Nietzsche seeks to transform the free will debate into an ethical discussion. In this view, the problem is about how thoughts and actions relate to oneself within one's own character, separating mere events from genuine actions, which is a feature only a moral agent who possesses free will can attain. The idea of self-relation has attracted considerable attention. Ken Gemes (2009), for instance, distinguishes two notions of free will. "Agency free will" is linked the specific conditions on genuine agency that constitutes an action as opposed to a doing that merely happens to one. "Desert" free will, which he thinks Nietzsche denies, relates to the idea that one deserves punishment. Similarly, Maudemarie Clark and David Dudrick (2009) argue that drives, according to Nietzsche, stand in an "order of rank" (*BGE* 6) based on normative authority. According to their interpretation, an agent's willing is constituted by this normative hierarchy of drives. When one acts in accordance with the authoritative drives and experiences the self-relation of commanding and obeying, the action is perceived as issuing from oneself—a view that appears to circumvent the constraints of determinism. However, these approaches remain entangled in the traditional framework of the free will debate, which seeks an independent precondition for moral agency and then examines its compatibility with determinism. This method encounters a fundamental problem: it reduces the discussion of moral responsibility and free will to an endless intellectual contest. Proponents on both sides construct elaborate conditions they claim that can define a moral agent, relying on rhetorical strategies to support their positions. Over time, this debate risks becoming a mere intellectual game, thereby forfeiting its practical relevance and moving further away from the lived realities of moral practice.

In this regard, P.F. Strawson stands out as an exception. His immensely influential essay "Freedom and Resentment" (1962) inspired what might be characterised as a paradigm shift in how we think about moral responsibility. Rather than identifying preconditions for moral agency or engaging in metaphysical debates about determinism, Strawson turns to our everyday moral practices and focuses on the link between reactive emotions and holding others responsible. As Strawson himself explains, "The object of these commonplaces is to try to keep before our minds something it is easy to forget when we are engaged in philosophy, especially in our cool, contemporary style, viz. what it is actually like to be involved in ordinary interpersonal relationships, ranging from the most intimate to the most casual" (1962). Hence, Strawson's strategy to find what is missing from the traditional debate is to turn away from the conceptual analysis of "freedom", "responsibility" and their preconditions. Instead, he focuses on "that complicated web of attitudes and feelings which form moral life as we know it" (Strawson, 2008: 34). What is essential is "the very great importance that we attach to the attitudes and intentions towards us of other human beings, and the great extent to which our personal feelings and reactions depend upon, or involve, our beliefs about these attitudes and intentions" (Ibid.). In other words, Strawson shifts the focus from abstract theoretical constructions of philosophical conceptions to the lived realities of our moral practices. For him, it is the interplay of our reactive attitudes towards others' behaviours and attitudes—such as resentment, gratitude, and forgiveness—that forms the essence of moral responsibility, rather than the abstract satisfaction of metaphysical preconditions. This perspective offers a compelling alternative to the traditional debate, rooting moral responsibility in human relationships and the shared experience of moral life as a natural phenomenon.

Notably, I believe that prior to Strawson, Nietzsche had already emerged as an outlier in this respect. While in the early stages of his career, he still approached the issue from the traditional perspective—focusing on the relationship between moral responsibility, free will, and determinism—later in his thought, Nietzsche transcended this framework and opted to replace the free/unfree will dichotomy with a distinction between strong/weak will. In this dissertation, I contend that Nietzsche endorses certain tenets central to Strawson's alternative approach to moral responsibility. My interpretation is grounded in the second essay of Nietzsche's *Genealogy of Morality* (*GM* II), in which he provides a genealogy of conscience, guilt, and responsibility, referring to it as "the long history of the origins of responsibility" (*GM* II: 2). Furthermore, Nietzsche praises individuals capable of certain moral practices, such as autonomy, value creation and taking responsibility, thereby demonstrating his practical and naturalistic understanding of moral responsibility. However, Nietzsche's remarks on these qualities remain underdeveloped: he does not provide a clear account of what these qualities mean, nor of how they differ from traditional metaphysical accounts of freedom and moral responsibility, which he elsewhere rejects. This lack of clarity leaves us with a set of normative ideals whose psychological structure is obscure, and here Strawson's framework proves illuminating. While Nietzsche does not offer a metaphysical theory of responsibility, he does present a rich but under-theorised account of what it means to be *responsible* in a positive, non-Christian sense. P. F. Strawson's sentimentalist framework is helpful for us to illuminate the emotional and psychological dynamics that underpin the positive stance of taking responsibility, as is exemplified by the sovereign individual. In particular, I aim to show how Nietzsche's conception of responsibility can be interpreted in terms of *reactive attitudes preserved under conditions of autonomy*, which reveals both the developmental trajectory of moral responsibility and the internal logic that links its various stages in *GM* II.

Indeed, I contend that Nietzsche goes further than Strawson. While Strawson restricts reactive attitudes to the moral domain and grounds them in an expectation of goodwill, a premise strictly confined to moral relationships, Nietzsche describes *reactive affects* that encompass a broader range of emotional responses, including anger, revenge, resentment, and guilt. These affects are not limited to moral expectations but can arise from any perceived violation of expectations. Moreover, they are rooted in more profound mechanisms, such as "bad conscience" and "guilt," which emerge from the processes of human existence and socialisation. More significantly, while Strawson uses the concepts of moral responsibility and free will interchangeably, Nietzsche distinguishes between them and, drawing on his moral views, offers a *practical and naturalistic understanding of free will*, which, I argue, provides a new perspective that transcends the traditional framework of contemporary debates on free will and moral responsibility.

If the interpretation of Nietzsche I propose in this dissertation is correct, he shares certain fundamental beliefs with Strawson's naturalistic approach to moral responsibility. These shared beliefs can be summarised as follows:

1. Responsibility, particularly moral responsibility, is grounded in moral practices such as praise, blame, and punishment. These practices reflect how we engage with each other in our shared moral life.
2. Those moral practices are constituted by our moral expectations of others and ourselves and their corresponding attitudes or emotional responses to these expectations. For Strawson, these are the "reactive attitudes," such as gratitude, resentment, and forgiveness. For Nietzsche, they are the broader "reactive affects," which include emotions like anger, ressentiment, revenge, and guilt.
3. Based on points 1 and 2, the metaphysical analysis of the conception of free will is unnecessary for defining whether an agent is morally responsible, because being an appropriate target of moral attitudes is not an issue of meeting certain predetermined conditions but about competence, depending on whether the agent is strong enough to hold oneself responsible within moral practices.
4. Hence, free will is not a prerequisite for being a moral agent; rather, it is a practical state achieved when taking on moral responsibility. The development of moral responsibility inherently cultivates the qualities necessary for free will.

If, as I will argue in this dissertation, moral responsibility and free will are forms of competence, then this aligns with Nietzsche's replacement of the dichotomy between free and unfree with that of strong and weak. This also clarifies the apparent contradictions in Nietzsche's texts. Sometimes, Nietzsche suggests that our common conceptions of free will are poisonous illusions. He points out that the agent is "fixed in the illusion of free will" (*HH* I: 106). Yet elsewhere, his practical philosophy appears to utilise this concept. In the second essay of *On the Genealogy of Morals*, Nietzsche praises the sovereign individual as the representation of "individual great men"(*UM* Ⅲ: 6) and ascribes "the consciousness of this rare freedom" (*GM* Ⅱ: 2) to this figure. As I will argue, moral responsibility is best understood as a form of competence, not dependent on traditional notions of free will. Becoming a moral agent does not require the ability to have done otherwise or being the ultimate source of one's actions. Instead, it involves possessing a certain kind of conscience. Thus, Nietzsche rejects the traditional metaphysical conception of free will while advocating for a form of free will that emerges in practice—a practical and naturalised one.

In the first two chapters, I will elaborate on these claims. I aim to demonstrate that many interpretations of Nietzsche's views on moral responsibility and free will from the traditional perspective are misguided. While Nietzsche initially adhered to the traditional framework of linking moral responsibility with the metaphysical preconditions of free will, his later work marks a departure. Like Strawson, he shifted the focus to moral practices rather than the prior conditions required to constitute moral agents. Specifically, In the first chapter, I plan to demonstrate that Leiter, a prominent interpreter who denies Nietzsche's concept of free will, mistakenly interprets Nietzsche's denial of traditional free will as rejecting moral responsibility, when in fact Nietzsche does not interpret free will and responsibility in this aspect. I will first consider whether the will has a causal effect that is required to give style to one's character (*GS* 290). Since Leiter uses experimental data from Libet (1985) and Wegner (2002) to support his interpretation of the epiphenomenon of the will, it will be productive to include a discussion of the work of experimental scientists (Schurger and Dehaene, 2012) to show that conscious thought plays a causal role in producing behaviour and shaping one's inner character. Then, to investigate whether Nietzsche is, as Leiter thinks, a hard-determinist, I will first focus on Nietzsche's texts in chronological order and investigate how his attitude toward the relation between determinism and free will evolves. I will investigate why Nietzsche, in his later work, goes so far as to reject the separation of free/unfree wills. If proposing strong and weak wills instead is a more adequate way to approach free will, it may be necessary to explore whether and how this notion connects with other conceptions beyond its traditional understanding..

In the second chapter, I argue that Nietzsche does not explore the connection between morality, free will and determinism. Instead, he perceives morality as a natural phenomenon that evolved alongside moral practices, which, in turn, contributes to free will as a form of strong will. To ensure that this interpretation is appropriate, I examine Nietzsche's claim that *GM* II presents "the long history of how responsibility originated" (*GM* II: 2) from a Strawsonian lens and argue that Nietzsche views moralised guilt as an unhealthy reactive response to inescapable guilt emotions. To accomplish this, I first unpack the notions of guilt as they build upon former conceptions and investigate how conscience, as a corresponding attitude toward breaking the rules, evolves into guilt and eventually incorporates self-worth concerning human character into its framework. Then I analyse Nietzsche's reasoning for rejecting free will by asserting that he believes free will is a fabrication by priests as a necessary precondition for the problematic Christian morality, which views individuals' personality as inherently evil, rendering their debts to God inexpiable and depriving them of the opportunity to alleviate guilt. In contrast, the healthier sense of guilt also engages self-worth, but focuses solely on what one has done. Therefore, Nietzsche does not argue against moral responsibility, but advocates for a healthy notion stemming from individuals' strength, which allows them to hold themselves responsible based on their pride and confidence to repay their misdeeds.

This raises an important question. Nietzsche suggests that, while the ancient Greeks exhibited the courage to take responsibility for their actions, he seems not to advocate for returning to the past. Instead, he praises figures who belong to the future, such as the sovereign individual, for their ability to create their own values. Since both ancient Greeks and the sovereign individual developed their version of conscience within the framework of the morality of custom (Ibid.) by obeying rules, it is crucial to examine how the social constraints shape an individual capable of creating her own values, thereby rendering the experience of guilt appropriate. To explore how to create values, I begin by considering how value creation is possible. I focus on Nietzsche's meta-ethical position on value by highlighting the similarities between value and taste and argue that similar to taste, the assessment of value is intrinsic to and constitutes the activity of valuing itself, which reconciles the subjectivity of value creation with the objectivity of value judgement. Then, to clarify the type of value Nietzsche advocates for and how to create values on this basis, I argue that good values are those, like good tastes, that are beneficial for unity and contribute to the development of further values. Nietzsche's proposed value creation, as I will argue, serves as a practical experiment to determine whether a value meets the demand for further development.

In the final chapter, I argue that while Nietzsche's affirmation of moral responsibility does not imply an affirmation of free will, he explicitly presents his own definition of free will as "the will to self-responsibility" (*TI*, Skirmishes, 39). Though free will and responsibility remain closely interconnected, free will is no longer understood as a mere capacity to choose or control, but rather as a state of perfection achieved through the possession of certain qualities. I propose that the elements I have identified in the development of conscience and the creation of values constitute Nietzsche's positive understanding of free will. Central to them is autonomy, signifying one's ability to maintain independence and forge one's own values. Additionally, creating values expressive of one's true character is grounded in one's knowledge of nature and self. Finally, Nietzsche closely links his ideal of freedom with necessity and fate. As Nietzsche emphasises, true freedom lies not in whether one's will is influenced by external circumstances, but rather on the affirmation of one's own worth and the acceptance of responsibility for one's own expectations.

# Chapter I

# Is Nietzsche a Sceptic on Free Will?

## 1. Introduction

It may be essential first to define the concept of free will in the traditional free will debate in order to avoid what Chalmers calls "merely verbal disputes" (Chalmers, 2011). In contemporary literature, free will is traditionally conceived as a kind of power to control one's actions in two different senses. The first sense of free will refers to a person's ability to do otherwise than she actually does. In this situation, the future is open and the agent has alternative choices other than what she actually did. Thus, this approach to free will is referred to as the "alternative possibilities" (AP) sense of free will. Another notion of free will does not focus on whether the future is open. For an agent to be free, her choices must originate from her control, meaning that she is the ultimate source of her actions. This kind of free will is referred to as "sourcehood conception" sense of free will. Robert Kane summarises these conceptions as follows:

> We believe we have free will when we view ourselves as agents capable of influencing the world in various ways. Open alternatives, or alternative possibilities, seem to lie before us. We reason and deliberate among them and choose. We feel (1) it is 'up to us' what we choose and how we act; and this means we could have chosen or acted otherwise. As Aristotle noted: when acting is 'up to us,' so is not acting. This 'up-to-us-ness' also suggests (2) the ultimate control of our actions lies in us and not outside us in factors beyond our control. (Kane, 2005: 6)

Brian Leiter interprets Nietzsche's free will as the power of control in these two traditional senses and focuses on Nietzsche's argument about the relationship between determinism and the causal role the will plays in bringing about actions. In his book, *Moral Psychology with Nietzsche* (2019), Leiter sums up his arguments in his early papers and claims that Nietzsche cannot affirm any form of responsibility and free will because Nietzsche believes in a distinctive kind of fatalism, which leads to his scepticism about moral responsibility and free will. If what one becomes and what one does is determined by forces beyond one's control, then no one should be "responsible any longer" (*TI* VI: 8). According to Leiter, Nietzsche's fatalism is based on his claim of "Doctrine of Types" (Leiter, 2002: 91), which posits that "each person has a fixed-psycho-physical constitution" (ibid.), defining them as a specific type of person. On this account, Nietzsche holds that every individual possesses an fated "essence" (*TI* VI: 8), making one a component of the whole necessity. A consequence of this is that one appears analogous to a plant whose development and possible life trajectories are determined by these constitutive "natural" facts or "type-facts" because they are "facts about the person's unconscious drives or affects‖" (Leiter, 2002: 91) and severely constrain one's future from the start.

Based on his diagnosis of fatalism, Leiter offers two possible models of willing to argue against the thesis that Nietzsche holds any version of free will: one is the 'epiphenomenal model' (Leiter, 2007: 13), and the other is the 'secondary cause model' (ibid.). According to the epiphenomenal model, our experience of exercising causal will is just an illusion because the experience of willing and the resulting actions are both products of type-facts, with no causal link between them. We mistakenly infer that the conscious will generates subsequent actions, but this ruling thought is actually determined by something other than the will. Nietzsche, in Leiter's view, believes that "conscious states are simply effects of underlying type-facts about [a] person" (Leiter, 2002: 92). Consequently, the conscious mental states play no causal role within the chain of events that produce actions. The secondary cause model, by contrast, posits that an external factor causes the experience of willing, and the will subsequently triggers the resulting actions. However, even if conscious mental states do play a part in the causal chain bringing about actions, they are not the ultimate cause of an action, thereby precluding the 'sourcehood conception' of free will.  A type-fact unconsciously "causes the experience of willing and then the will causes the action" (Leiter, 2007: 13). In this way, the will's causal efficacy in producing actions remains itself an outcome of unconscious determinants, which is not strong enough to support attributions of moral responsibility and free will.

Leiter thinks that it is hard to tell which model of willing Nietzsche takes because Nietzsche's texts are intrinsically ambiguous regarding his stance on the will, but he is inclined to read Nietzsche as committed to the epiphenomenal model, which he thinks is "(a) supported by his texts and (b) most likely to be correct as a matter of empirical science" (Leiter, 2007: 13). His first strategy is to defend his epiphenomenalism with two arguments: (1) confusing cause and effect, regarding the illusory experience of willing, and (2) the real genesis of action (Leiter, 2019), which concerns the way actions are brought about. He then turns to empirical findings from Libet and Wegner to assess the causal structure of agency as Nietzsche describes it. If Libet and Wegner successfully prove that the initiations of actions are unconscious factors rather than the will, that will be a good reason to read Nietzsche as holding the epiphenomenal model and rejecting free will. However, there are several challenges to Leiter's interpretation. In this chapter, I will first argue against the epiphenomenal model from the following three aspects, (1) the text that Leiter relies on support the epiphenomenonal interpretation does not definitively rule out the possibility that consciousness play a causal role; (2) his epiphenomenonal interpretation of recent psychological experiments is unwarranted, as alternative explanations of the data remain viable; (3) there are some empirical studies designed to assess the influence of conscious thought on behaviour. Following this critique, I will argue that the remaining secondary cause model leaves some room for free will. In his discussion about the phenomenology of willing, Leiter acknowledges that it is still possible for the experience of willing to mislead "us as to our real will, and that this real will does, in fact, stand in the appropriate causal relationship with actions" (Leiter 2007). Based on Nietzsche's texts, I claim that Nietzsche is not, as Leiter thinks, an incompatibilist who takes *causa sui* as a prerequisite for free will. Rather, Nietzsche not only rejects the notion of *causa sui* but also changes his views dynamically from his middle to late periods. Moreover, he distinguishes his conception of free will (strong will) from other formulations, which suggests that his rejection of a certain type of free will does not necessarily make him a sceptic about moral responsibility and free will as a whole.

## 2. The Epiphenomenon of Will

### 2.1 Confusing Cause and Effect

As I mentioned above, Nietzsche, in Leiter's view, endorses the notion of type-facts and maintains that each individual possesses an immutable essence that determines who she is. What governs one's thoughts and behaviours is not conscious will but rather something else beneath consciousness. However, as Nietzsche writes in *Daybreak*, regardless of whether unconscious factors determine human behaviour, people tend to adopt an agnostic attitude towards them, as they are "accustomed to exclude all unconscious processes from the accounting and to reflect on the preparation for an act only to the extent that it is conscious" (*D* 129). The result is that errors arise concerning the relationship between consciousness and actions. To prove his argument, Leiter draws on "The Four Great Errors" section of *Twilight of the Idols* to support his epiphenomenalism interpretation of the will and concludes that Nietzsche is sceptical of the causal efficacy of conscious will. This section outlines three errors related to causation: "the error of confusing cause and effect" (1-2), "the error of false causality (3), and "the error of imaginary causes" (4-6), all of which culminate in the last error, "error of free will". To fully grasp Leiter's argument, it is crucial to reconstruct the reasoning process that connects the first three errors to the conclusion that free will is an illusion.

The first error is "confusing cause and effect", which means that we tend to confuse the cause and effect when we overlook their mutual "deep cause" and instead rely on regular correlations. In Nietzsche's example of Cornaro, the Italian mistakenly regards his diet as the condition for a long life, while what explains his diet and long life is the slowness of his metabolism. Cornaro's advocacy of a particular diet for longevity mirrors the way morality and religion prescribe formulas for human behaviour. They persuade that "a well-turned-out human being, a 'happy one,' must perform certain actions and shrinks instinctively from other actions" (*TI* VI: 1). Nevertheless, just as Cornaro is not free to eat more or less, we are not free to determine our actions, which are dictated by physiological factors.

This leads to the second error, "false causality", which refers to "the mistake of thinking we know what causation is because of our introspective confidence in what we take to be the causal powers of our own mental life" (Leiter, 2019: 134). In this section, Nietzsche argues that the causal powers we attribute to our will are merely illusions we impose upon ourselves. In fact, it is not an ego that causes actions. "The so-called motive: another error" does not explain anything, but "merely accompanies events" (*TI* VI: 3). For Leiter, although Nietzsche does not present a formal argument for false causality but merely asserts it, his belief in type-facts is clearly at work here. Although these type-facts operate below the level of consciousness, they play a major role in contributing to consciousness in ways that people always ignore. To be more specific, if conscious thoughts, following actions, and the propositional content that makes them appear to be causally connected are all the products of type-facts, then they are in fact epiphenomenal and causally inert —meaning they are by-products of type-facts without any causal power of their own. Even if consciousness possesses causal power, it remains governed by unconscious factors. In this picture, type-facts are seen as the real drivers behind human behaviour, rendering conscious thoughts merely surface phenomena. Thus, while people may believe that their conscious decisions lead to their actions, Nietzsche's view, as interpreted by Leiter, suggests that this belief is an illusion—an instance of mistakenly attributing causal efficacy to consciousness. The second error thus challenges the notion of conscious will by highlighting the predominant influence of unconscious factors.

The third error, "the error of imaginary causes", arises when we experience a feeling and "we are never satisfied merely to state the fact that we feel this way or that: we admit this fact only —become conscious of it only—when we have furnished some kind of motivation" (*TI* VI: 4). Nietzsche claims that we seek justification for this feeling and we become conscious of it only when we have invented a cause for the way we feel as we do. This tendency stems from the fact that, when faced with fear, individuals instinctively seek preferred and habitual explanations. "To derive something unknown from something familiar relieves, comforts, and satisfies, besides giving a feeling of power" (Ibid.). For example, as Nietzsche puts it, the Christian will think of "sin". In fact, "the whole realm of morality and religion belongs under this concept of imaginary causes" (Ibid.). The occurrence of agreeable feelings is attributed to faith in God. An individual "is in a state of hope" and experiences inner peace by adhering to Christian morality and refraining from "ill-considered actions" (Ibid.). One does so because she translates her pleasurable feelings into a false imagination and hopes to get a rest by believing in God. Most people are indifferent to the real causes of thoughts and actions; rather, they require the false knowledge of causes, which is "a bed to rest on, or the way to such a bed, or a diversion, or a form of leisure" (*GS* 324). The error of imaginary causes is just another manifestation of the confusion between cause and effect, since when one imagines an effect as the cause for something else, one overlooks that both of them are products of a deeper underlying cause, thereby perpetuating the confusion.

As a consequence of the three preceding errors concerning causality, the last error is "the error of free will." For Nietzsche, the doctrine of the will, which renders humanity "accountable", "has been invented essentially for the purpose of punishment, that is of finding guilty" (*TI* VI: 7). In this sense, morality and religion rest upon a flawed conception of agency, as individuals are presumed to be free and thus accountable, despite that they could not have done things other than they are determined to do. To be free and morally responsible requires the will to be causal, but the three errors concerning causality suggest that our experience of willing is merely an illusion.

Nevertheless, contrary to Leiter's conclusions, these errors in causality do not rule out the possibility that conscious states play a role in the causal chain. They merely indicate that people sometimes make mistakes in believing in the causal efficacy of conscious states when it does not have, but do not mean that people always make such mistakes. For example, an agent edified by Christian morality is accustomed to taking the feeling of hope and obeying rules as the cause for pleasurable feelings. However, she is merely misled into giving up her power of control, instead of succumbing to the illusion woven by Christianity. Her misunderstanding of causal relation in the etiology of action does not negate the possibility that she retains the capacity to will causally. Just as making an error in solving a mathematical problem does not imply that no correct solution exists, the presence of errors in causal reasoning does not necessitate the conclusion that consciousness lacks all causal power. It is true that in certain cases, individuals invent imaginary causes of actions and get a feeling of command, when in reality, unconscious factors are the true drivers of behaviour. Yet, this error is merely a misunderstanding of the logical sequence between consciousness, unconsciousness, and behaviour, rather than demonstrating that consciousness cannot exert causal influence at all in any situation. Therefore, it is still difficult to make the claim that a type-fact is the only deep cause that brings about actions and that consciousness is universally devoid of causal influence. The argument concerning errors in causality leaves room for the possibility that the conscious will still retains causal efficacy. Since Leiter recognises this, Leiter advances his secondary cause model. We will return to this topic later.

### 2.2 The Real Genesis of Action

To follow Leiter's error argument, if the experience of willing does not genuinely explain how actions are brought about, then what plays a causal role in the chain should be something other than the will. As previously discussed, the argument concerning type-facts, which refer to psycho-physical "facts about the unconscious psychology and the physiology of agents" (Leiter, 2019: 141), is the foundation for the errors of causality. Because people are always misled into correlating conscious states with actions, they take it for granted that conscious state plays a causal role in bringing about actions. However, according to Leiter, what really happens is that unconscious type-facts account for both our experience of willing and our actions.

From the argument Leiter proposes, we can see that Leiter is prone to the stronger claim about the epiphenomenal status of conscious willing. However, merely reading the texts Leiter cites is insufficient to conclude that Nietzsche endorses this view, as the causality errors still allow for the possibility of a causal will. Leiter admits that Nietzsche's texts are ambiguous between which of the two models he embraces. Therefore, he turns to empirical research and claims that when deciding which model we ascribe to Nietzsche, we should find a view that is both supported by (a) Nietzsche's texts and (b) the result of empirical science. If empirical science indeed confirms that conscious will lacks causal efficacy, and Nietzsche's texts can accommodate both models, then it would be reasonable for us to follow Leiter and prioritise the epiphenomenal model as the primary one Nietzsche defends. To support his interpretation, Leiter appeals to empirical research by Benjamin Libet and Daniel Wegner, both of whom challenge the will's role in causal reality. However, I argue that the empirical research Leiter reviews is so controversial that it fails to provide unequivocal support for the epiphenomenal model. I then turn to the work of Meles and Schurger and argue that Leiter's epiphenomenonal interpretation of recent psychological experiments is unwarranted. Furthermore, experiments designed to measure the effectiveness of conscious thoughts on actions, which is synthesised by Baumeister, Masicampo and Vohs (2011) provide empirical support to the claim that the will has causal efficiency.

### 2.2.1 Empirical Research

(a). Libet and Wegner

The first well-known neuroscientific study examining free will was conducted by Libet (1983), focusing on how unconscious intentions affect allegedly free and voluntary decisions. Libet's work drew upon Kornhuber and Deeck's (1965) discovery of the Bereitschaftpotential (readiness potential, or RP), described as "the electro-physiological sign of planning, preparation, and initiation of volitional acts" (Kornhuber & Deecke, 1990). If conscious states do have causal efficiency, the subject ought to be aware of the intention to act prior to the onset of RP. However, the result of Libet's experiment contradicts this prediction. In his experiment, Libet asked the participants to move their right wrist whenever they wanted to, but to keep track and report the precise time when they had the intention to do so, with the help of a clock in front of them. Libet found that RP emerged unconsciously approximately 550 ms before the action occurred. However, his participants reported that they became aware of the intention to act only 200 ms before the wrist flexion. There was a 350 ms temporal gap between the urge and RP. In this sense, "the brain starts doing something first…then the person becomes conscious of wanting to do the action" (Wegner, 2002: 53). It seems implausible that the volition process could be initiated by awareness of these actions, given that it arises even later. Thus, Wegner sums up his experiment that "free will and free choice of whether *to act now* could not be the initiating agent, contrary to one widely held view" (Wegner, 2002: 54). As for Wegner, the phenomenology of willing and the causation admittedly come apart. "The automatisms and illusions of control… remind us that action and the feeling of doing are not locked together inevitably. They come apart often enough to make one wonder whether they may be produced by separate systems in the mind. The processes of mind that produce the experience of will may be quite distinct from the processes of mind that produce the action itself" (Wegner, 2002: 11). As a result, the real genesis of action should be attributed to unconscious causes, suggesting that both conscious willing and action are the products of unconscious facts. "Unconscious and inscrutable mechanisms create both conscious thought about action and the action, and also produce the sense of will we experience by perceiving the thought as cause of the action" (Wegner, 2002: 98). If Libet and Wegner are correct, this suggests that a century after Nietzsche, empirical psychology has produced evidence supporting the epiphenomenonal reading of willing that Leiter attributes to Nietzsche.

However, the interpretation of Libet's empirical research is not conclusive and should be examined in more detail. Some philosophers, like Mele and Nahmias, pose several questions to this research. First, both argue that a distinction should be made between proximal and distal intentions (Mele, 2006; Nahmias, 2010). Distal intentions concern future actions, whereas proximal intentions pertain to immediate actions. The problem is that in Libet's experiment, the effect of distal intention was ignored and not tested at all. Libet conflated proximal and distal intentions and carried out the experiment without realising the possibility of further differentiation. Thus, even if Libet's experimental findings are valid, they merely demonstrate that conscious thought lacks causal efficacy in initiating actions linked to proximal intentions, while it may play a role in influencing actions associated with distal intentions.

Furthermore, Nahmias (2010) argues that Libet's findings are compatible with two alternative interpretations that do not negate the efficacy of conscious will. The first one is that while the unconscious urge (RP) may initiate the beginning of the action, conscious will can still influence whether to act on this impulse and determine the manner of its execution. Indeed, Libet himself acknowledged that "preparatory cerebral processes can be blocked consciously just prior to their consummation in actual motor outflow" (Libet, 1985: 538). In *The Gay Science*, Nietzsche also expresses a similar idea. He distinguishes "the cause of acting" from "the cause of acting in a particular way, in a particular direction, with a particular goal" (*GS* 360). The cause of acting "is a quantum of dammed-up energy that is waiting to be used up somehow, for something", while the second kind of cause includes "purpose". Nietzsche here appears to suggest that people habitually conflate these two causes, mistakenly perceiving purpose as the primary driving force behind actions, whereas it merely serves as a directing force. The upshot from this passage is that although there may be errors of confusing cause and effect, and even if the unconscious urge primarily drives actions, conscious thought still plays a crucial role in directing and shaping their expression. Some might question whether a directing force is strong enough to be regarded as a cause. In addressing this problem, it is helpful to refer to Mackie's notion of INUS conditions (an Insufficient but necessary part of an Unnecessary but Sufficient condition) (Mackie, 1965: 245-264). Within this framework, conscious will alone, which acts as a directing force, is insufficient by itself to cause an action but remains a necessary component of the causal chain, because the specific direction of actions will not occur even if other factors are present without conscious will. This specific causal chain, which incorporates conscious factors, is unnecessary in the sense that it is not the only possible way by which actions can occur; however, when combined with other factors, conscious will forms a sufficient condition and creates a complete causal pathway leading to the actions. Thus, conscious will remains an integral component of the causal process, demonstrating that the significance of the unconscious urge does not preclude the causal efficacy of conscious will.

Another possible interpretation does not centre on whether conscious thoughts are causal or not in the process initiated by unconscious urges, but instead contends that "early non-conscious brain activity just is part of the correlate of the conscious intention, or part of the necessary build-up to such intentions or decisions" (Nahmias, 2010: 352). According to this interpretation, certain actions may be accounted for by an unconscious process; however, unconscious urges themselves originate from conscious intentions, which, in turn, may facilitate the emergence of later conscious processes. As Nahmias posits, conscious thoughts can be caused by "earlier complexes of events, including external stimuli and neural activity, some of which may have been caused by even earlier conscious processes" (ibid.). In Libet's research, participants were provided with instructions by Libet and subsequently formed their distal intentions upon these instructions. Subsequently, they became aware of the conscious urges. In this situation, their awareness of the urges is triggered by unconscious intentions, but certain complexes of psychological factors, including conscious thoughts, may precede these unconscious intentions. For example, they deliberately instructed themselves to comply with the experiment's rules in order to secure a reward, thereby remaining observant and responded to Libet's instructions immediately even in the absence of explicit conscious realisation (with their attention possibly directed towards movement rather than recording the timing of their intentions)!

The most difficult challenge to Libet's experiment comes from Aaron Schurger's question of the relation between RP and actions. He reviewed previous studies and found that when we make a decision based on visual experiences, assemblies of neurons begin to accumulate visual evidence in favour of diverse outcomes. In other words, a decision is initiated when one outcome is strong enough to cause its neuronal aggregation to exceed a threshold. However, Schurger and his colleagues argued that brain activity comprises merely random neuronal fluctuations rather than conscious outcomes. To prove this, they replicated Libet's experiment but introduced something different, which they called *Libetus Interruptus*. In Schurger's experiment, the majority of the task remains identical to the original, except for a click that signals the participants to act promptly upon hearing it. The research found that a rapid response to the click occurs when the accumulation of neural noise is near the threshold, which could be seen as RP in the electroencephalogram recordings, whereas in cases of slower responses, RP is absent. It appears that action is triggered when neuronal fluctuations accumulate sufficiently to exceed a certain threshold. Thus, Schurger challenges the causal role RP plays and thinks that RP may merely reflect stochastic fluctuations rather than a neural decision to move. "According to our model, uncued movements in a task like Libet's tend to be preceded by a gradual increase in neural activity [measured at the scalp or the single-neuron level] whose causal role is incidental—not directed (consciously or otherwise) at producing a movement" (Schurger, 2012: 6). If RP does not causally explain following actions, then the premise of Libet's conclusion, that is, that some antecedent precedes conscious intentions to cause actions, fails. Our brains do not decide for us prior to our realisation, because RP is merely a reflection of the ebb and flow of neural activity, which just happens to correspond to a decision afterwards. Although the conclusion of Schurger's experiment does not provide affirmative support for the existence of free will, it at least calls into question the reliability of Libet's argument.

Given that there are several possible interpretations and even direct challenges to Libet's experiment, it appears that the empirical research employed by Leiter cannot unequivocally support his epiphenomenonal model. Furthermore, alternative accounts of conscious thought remain open in contemporary empirical research. A growing amount of experimental evidence supports the view that conscious thought plays a causal role, even if only as a secondary cause, which reinforces the argument that Nietzsche should not be construed as endorsing the epiphenomenonal model that outright denies the causal efficacy of consciousness. While Nietzsche's texts do not definitively favour one model over another, the secondary-cause model offers a plausible interpretation, which, in addition, finds further empirical support in scientific experiments.

(b). The Causal Efficacy of Consciousness

In "Do Conscious Thoughts Cause Behavior?," Baumeister, Masicampo and Vohs examined hundreds of recent experiments and concluded that "the evidence for conscious causation of behaviour is profound, extensive, adaptive, multifaceted and empirically strong" (Baumeister, Masicampo, & Vohs, 2011: 331). They identified four broad dimensions through which consciousness influences behaviour. Firstly, consciousness is useful for replaying past events, reflecting on previous performances, inferring implications of recent occurrences, and planning for future outcomes, thereby enabling current behaviour to benefit from prior events and to shape future ones (e.g., Anseel et al. 2009). The following examples are representative:

> Anseel et al. (2009) provided participants with task feedback on a web-based work simulation task. Some were taught to go back and review their thoughts and actions, while feedback led to significant improvements in later performance. Reflection without feedback brought no benefit, as others have also concluded (see Mayer 2004, on the uselessness of reflection without external guidance or feedback). (Baumeister, Masicampo, & Vohs, 2011: 338)

> Epstude & Roese (2008) provided an overview and theoretical integration of how counterfactual replays affect later behaviour. They provided evidence that counterfactual replays have two sorts of effects. First, they stimulate specific intentions to behave differently in similar situations subsequently, and these intentions do influence subsequent behaviour. Second, they can have more general, content-neutral effects such as by altering mindsets and motivational states, which can then affect behaviour. Roese (1994) showed that after an initial anagram task, engaging in upward counterfactual replays caused improvements in subsequent performance, whereas downward counterfactual replays did not improve performance relative to neutral control. (Baumeister, Masicampo, & Vohs, 2011: 339)

To be more specific, even though conscious elements alone are not the direct causes of these actions, their replay and reflection on past events, as well as future planning, can be seen as parts of a broader causal mechanism that encompasses both conscious and unconscious elements. In this way, consciousness qualifies as causal because it actively influences and directs behaviours through learning, evaluation, adaptation, and planning.

Secondly, consciousness allows an individual's behaviour to be informed by social and cultural events. It facilitates the sharing of information with others, the understanding of others' perspectives, and the management of social norms (e.g., Fazio et al. 2004). For example, Fazio et al. (2004) instructed participants to play a game in which they chose which beans to eat, only some of which reinforced the eating by providing valuable energy points. Participants quickly learned to eat only the helpful ones. However, when the experimenter provided bogus tips regarding which beans were advantageous, participants began consuming those while avoiding the other beans (Baumeister, Masicampo, & Vohs, 2011: 348). Another demonstration by Jorgenson & Papciak (1981) involved participants playing a game in which mutual restraint was necessary for long-term gain. The results indicate that communication can improve group performance.

Experimental evidence is also plentiful in support of the conclusion that conscious thoughts are beneficial in situations where individuals are confronted with multiple alternative possibilities (e.g., Zeelenberg et al. 1996). In an experiment summarised by Zeelenberg and Pieters, participants were required to choose between a safe and a risky gamble. The results showed that even when participants did not select a particular option, knowing the outcome of that gamble influenced their decision-making. This tendency was driven by the desire to avoid the regret of realising that an alternative choice might have yielded a better result (Zeelenberg et al. 1996). Baumeister and his colleagues summarised the discussion by asserting that "Conscious thought can simulate alternative realities and by imaging them increase the likelihood that they will come true … A vital function of consciousness may be to comprehend the multiplicity of possibilities so as to facilitate bringing about a preferable one" (Baumeister, Masicampo, & Vohs, 2011: 353).

Finally, the authors also found that virtually all human behaviour emerges from the complex interaction of conscious and unconscious states. They summarised thus: "Nothing we have reviewed would prove that any behaviour emerged from exclusively conscious processes. Likewise, ostensible evidence of unconscious causation is typically compromised by extensive reliance on conscious processes too, such as for giving instructions and focusing attention; the participant is merely unconscious of one particular link in the causal chain" (353). It is worth noting that this aspect of the complex interaction between consciousness and unconscious states corresponds closely with Nietzsche's account of consciousness. As mentioned above, the distinction between "the cause of acting" and "the cause of acting in a particular way, in a particular direction, with a particular goal" (*GS* 360) suggests that while unconscious drives may initiate actions, consciousness, as an integral component of the causal process, helps shape and direct how actions are executed. The other three aspects—replaying past events, being informed by social and cultural events and aiding decision-making in the face of multiple possibilities—can all be regarded as outcomes of the dynamic interplay. For example, by saying that consciousness enables one's behaviour to be informed by social and cultural events does not imply that consciousness, in isolation, can actively get information from social norms and respond accordingly. A conceivable scenario is that an individual might unconsciously absorb social norms through repeated exposure, yet consciously reflect upon these norms to decide how to act in a specific situation.

Therefore, numerous studies, including those presented above, support the conclusion that conscious thought indeed plays a causal role in bringing about actions, even if it is not the ultimate cause but merely a secondary cause, which is itself initiated by unconscious elements. The experimental evidence employed by Leiter to support his reading of Nietzsche is undermined by alternative interpretations offered by Nahmias and Mele. Moreover, further scientific experiments make the secondary-cause model not only align with Nietzsche's texts but more likely to be correct from an empirical standpoint. However, even if the view that the will functions as a secondary cause is the one Nietzsche upholds, incompatibilism still poses a threat to the possibility that Nietzsche can hold any kind of free will. In the next section, I will discuss Leiter's argument for incompatibilism and argue that Nietzsche is not an incompatibilist.

## 3. The Incompatibilism Argument

In the previous section, I concluded that since Leiter's interpretation of the epiphenomenal model cannot be simultaneously supported by both Nietzsche's texts and experimental evidence, we should provisionally interpret Nietzsche as holding the view that conscious will does indeed cause actions, even if only as a secondary cause. However, Leiter has another resource in attributing an incompatibilist position to Nietzsche and contends that if Nietzsche maintains this view, then according to Nietzsche, no one can act freely or take responsibility for their actions. Leiter's strategy is to examine the constitution of willing in order to scrutinise the experience of exercising control, a phenomenon he terms the "phenomenology of willing" (Leiter, 2019).

### 3.1 The Phenomenology of Willing

Leiter regards the key discussion of the phenomenology of willing as appearing in Section 19 of *Beyond Good and Evil*. According to Nietzsche, every act of willing comprises three components. "A plurality of sensations" refers to the feelings of the body, while "a ruling thought" denotes the idea of controlling bodily feelings. In addition, "the affect of the command" (*BGE* 19) causes us to identify with our "ruling thought", which produces the feeling of superiority and leads us to recognise our identity as the commander during the process of self-identification. "In short, one experiences willing when one feels as if the bodily qualia are obeying the thought, and that the commandeering thought is who I am" (Leiter, 2019: 127). However, Leiter thinks that Nietzsche highlights a paradox here, namely, that we are "at the same time the commanding and the obeying parties" (*BGE* 19). The reason an individual identifies with the feelings of commandeering rather than obeying is that one believes in the will's success in bringing about actions and "enjoys an increase of the sensation of power which accompanies all success" (Ibid.). Nonetheless, the feeling of superiority attached to the commandeering thought is unrelated to the causal relationship between the will and actions. As Nietzsche puts it, "a thought comes when 'it' wishes, and not when 'I' wish" (*BGE* 17). Since it is not a subject that performs the thinking, it follows that a thought entering consciousness is not preceded by willing. In other words, the "commandeering thought" to which the feeling of superiority attaches is absent. If our thoughts are not willed in the manner we assume, then the actions that follow from our thoughts and the experience of willing are insufficient to underpin moral responsibility and free will.

Next Leiter claims that the feeling of superiority with which we identify does not, in fact, reflect an actual causal chain. Leiter's argument rests on two premises. The first premise is that the ruling thought itself is causally determined by something other than the will. The second premise is incompatibilism, which posits that being self-caused is a necessary condition for moral responsibility. As he writes, "the decisive component of the experience of will—namely, the commandeering thought—is, itself, causally determined, at some point, by something other than the will, since thoughts that come into consciousness are causally determined by something other than the will" (Leiter, 2019: 131). Since the ruling thought, as a component of the experience of will, is not *causa sui*, it follows that the will which it helps to constitute is not *causa sui*, and consequently, any actions that follow the phenomenology of willing cannot bear any moral responsibilities (Leiter, 2019). The incompatibilism argument can be formulated as follows:

(1) A person acts freely only if she is *causa sui*.

(2) No one can be *causa sui*.

(3) Therefore, no one can act freely.

It seems that even if the will does play a causal role in bringing about actions, its determination by factors beyond conscious processes places the ultimate source of actions beyond one's control, which rules out the possibility that Nietzsche has a positive notion of free will. To investigate if Nietzsche in fact holds the premises and the following conclusion that Leiter ascribes to him, I have to trace back to Nietzsche's texts in three of his work concerning his stance. I will show that Nietzsche's view on free will undergoes a dynamic evolution, which means that his perspective on free will shifted significantly. When writing *Human, All Too Human*, he hoped to achieve a state of free spirit but found that the free will at hand was an error designed by morality. This illusory notion of free will holds that people are accountable for their actions, but Nietzsche was a hard determinist and source-incompatibilist. According to source-incompatibilism, the world is determined such that we cannot be the ultimate cause of our actions, which indicates that no one can act freely and be responsible for her actions. This reading matches up to what Leiter ascribes to Nietzsche. However, Nietzsche doubted the existence of free will, while dreaming of a form of free will divorced from necessity, and his view on this matter changed during the time he wrote *Beyond Good and Evil*. In *BGE* 21, Nietzsche seems to question the concept of *causa sui*, and rejects determinism and the inference from "psychological necessity" to "unfree will". Rather, he introduces a new distinction between "strong" and "weak wills". The distinction between strong and weak wills is the one Nietzsche holds on to in *Twilight of the Idols* and associates with his notion of the "sovereign individual" in *On The Genealogy of Morals*, which is connected to his positive ascriptions of free will and moral responsibility.

### 3.1.1 *Human, All too Human* (1878)

Although Nietzsche regards *HH* as a work dedicated to the free spirit, he argues against free will in several passages. One of his strategies to refute free will is by asserting the necessity of all events. In section 106 of *HH*, he compares human actions to the waterfall and says that at the sight of a waterfall, we think we see countless curvings and the freedom of the will, "but everything here is necessary, every motion mathematically calculable" (*HH* 106). The same applies to human actions. "If one were all-knowing, one would be able to calculate every individual action, likewise every advance in knowledge, every error, every piece of wickedness" (Ibid.). Nietzsche also points out that the agent is "fixed in the illusion of free will". He holds the view that the state of the past with the law of nature makes every human behaviour in the future predictable. Given that the world we live in is deterministic, free will "is itself part of the mechanism it would have to compute" (Ibid.). In this section, Nietzsche explicitly states that determinism negates the possibility of free will.

In the next section, Nietzsche continues to say that we are living in necessity:

> All his evaluations, all his feelings of respect and antipathy have thereby become disvalued and false: his profoundest sentiment, which he accorded to the sufferer, the hero, rested upon an error; he may no longer praise, no longer censure, for it is absurd to praise and censure nature and necessity. (*HH* 107)

Furthermore, Nietzsche thinks that human actions are just like a fine work of art or a plant. Neither a work of art nor a plant possesses the capacity to do something for itself. Similarly, an agent living in necessity also lacks the capacity for the same thing. For Nietzsche, necessity and determinism render human actions unworthy of either praise or blame. The notion of accountability, Nietzsche argues, is a falsehood to which we cling out of confusion.

Nietzsche's stance on the relationship between free will and determinism is most explicitly articulated in section 39. In the section titled "The fable of intelligible freedom", Nietzsche asserts that free will is an error. Nietzsche begins his argument by tracing the origins of morality to the pleasurable or harmful consequence of human actions, but he then realises that "we forget the origin of these terms and imagine that the quality 'good' or 'evil' is inherent in the actions themselves, without consideration of their consequences" (*HH* 39). From holding each other accountable for the consequences of human actions, we soon move to holding each other accountable for the actions themselves. From holding each other accountable for the actions themselves, we move to holding each other accountable for the motives. Finally, this logic extends to holding individuals accountable not merely for their motives but for their very nature. However, one finally realises that human nature itself cannot be subject to accountability, as it is merely a product of necessity and "assembled from the elements and influence of things past and present" (*HH* 39), which makes one accountable for nothing. As a result, the process of reasoning from consequences to the will constitutes a history of error—one that has conditioned us to feel responsible when we have a bad conscience after certain actions. We rarely come to grasp that "the history of moral feelings is the history of an error, an error called 'responsibility', which in turn rests on an error called 'freedom of the will'" (Ibid.). In this paragraph, Nietzsche articulates the idea that an agent cannot be truly responsible for her actions, as these actions stem from motives and a nature shaped by past and present influences. To be genuinely responsible for something, the agent must be the ultimate source of her actions. If one cannot be the ultimate source of her own nature, she cannot be responsible for her motives, actions or their consequences. Nietzsche further argues that one can disaccustom oneself to the feeling of freedom. For Nietzsche, "this displeasure is a habit that can be given up…Tied to the development of custom and culture, it is a very changeable thing" (Ibid.). The only reason for the false belief in free will and moral responsibility is "from fear of the consequences" (Ibid.), which stems from the anxiety of accepting that one's actions are not a result of personal choice. In other words, one requires the illusion of free will because of a fear of the potential nihilism and the loss of moral accountability.

Some may wonder why Nietzsche, on the one hand, writes his book for free spirits, who possess freedom of the will and can freely determine actions, but on the other hand, rejects free will. Nietzsche offers an implicit answer in the first preface. As he writes, it is true that free spirits "do not exist, did not exist" (*HH* I: P: 2), but he still "had need of them" (Ibid.). He regards himself as one struggling to achieve free will as a means of confronting wickedness. Although it is clear that Nietzsche was a determinist and source-incompatibilist at the time he wrote *Human, All too Human*, holding the idea that no one can be free because determinism makes it impossible for one to be the ultimate source of her actions, these passages do not straightforwardly imply a complete rejection of free will. For Nietzsche, the kind of freedom presupposed by free will here is not yet "that mature freedom of spirit which is equally self-mastery and discipline of the heart and permits access to many and contradictory modes of though"(*HH* I: P: 4), so it can be rejected by necessity. However, he views himself as being on the path toward this mature freedom. "It is still a long road" (Ibid.), yet this journey ultimately reaches its conclusion in his later works.

### 3.1.2 *Beyond Good and Evil* (1886)

Nietzsche's idea about necessity dramatically changed when he wrote *Beyond Good and Evil*. In section 21, he claims that:

> The *causa sui* is the best self-contradiction that has been conceived so far, it is a sort of rape and perversion of logic; but the extravagant pride of man has managed to entangle itself profoundly and frightfully with just this nonsense. The desire for "freedom of the will" in the superlative metaphysical sense, which still holds sway, unfortunately, in the minds of the half-educated; the desire to bear the entire and ultimate responsibility for one's actions oneself, and to absolve God, the world, ancestors, chance, and society involves nothing less than to be precisely this *causa sui*. (*BGE* 21)

In this paragraph, Nietzsche maintains the view that being *cause sui* is the necessary condition for having a certain kind of free will and moral responsibility. This can be seen as a continuation of his view in *HH*, where he claims that to be responsible for one's deeds presupposes responsibility for one's own nature. However, two differences emerge here, which indicate Nietzsche no longer holds the same position during the time he wrote *BGE*. First, Nietzsche adds some restrictions to the concept of free will and responsibility. He rejects merely free will "in the superlative metaphysical sense" and "ultimate responsibility", which "absolve God, the world, ancestors, chance, and society" rather than referring to all forms of free will and responsibility. For Nietzsche, being *cause sui* is merely a necessary condition for a certain kind of free will and a certain kind of responsibility. Second, Nietzsche claims that *causa sui* is a self-contradictary concept and "is a sort of rape and perversion of logic" (*BGE* 21). There may be two reasons for Nietzsche's rejection of *causa sui*. One reason lies in the fact that to be *causa sui*, a cause has to bring itself into existence, which requires the cause to be logically prior to its own existence. However, being prior to its own existence is absurd and self-contradictory. Second, whether or not one is *causa sui*, both two states appear to rule out free will and moral responsibility. If one is determined, then alternative possibilities and the ultimate source of one's actions are beyond one's control. But in an indeterministic world, everything happens randomly and accidentally, for which no one can be held responsible. Therefore, this concept itself includes inherent self-contradiction.

Nietzsche goes on to argue that we should put the notions of "free will" and "non-free will" out of our heads, for the idea that there is a "non-free will" is the result of misuse of cause and effect. "One should not wrongly reify 'cause' and 'effect,' as the natural scientists do (and whoever, like them, now 'naturalises'[1] in his thinking)... one should use 'cause' and 'effect' only as pure concepts, that is to say, as conventional fictions for the purpose of designation and communication —not for explanation" (Ibid). For Nietzsche, "cause" and "effect" are useful fictions for us to communicate and share ideas, but they are merely pure concepts, which do not refer to any entity in the world. This leads to the rejection of necessity and the notion of non-freedom:

> In the "in-itself"[2] there is nothing of "causal connections," of "necessity," or of "psychological non-freedom"; there the effect does *not* follow the cause, there is no rule of "law." It is *we* alone who have devised cause, sequence, for-each-other, relativity, constraint, number, law, freedom, motive, and purpose; and when we project and mix this symbol world into things as if it existed "in itself," we act once more as we have always *acted*—

[1] This is a different sort of naturalising than the one I support and attribute to Nietzsche later in this dissertation. "Naturalise" here refers specifically to how natural scientists treat conceptions—by regarding them as independent existence.

[2] By saying "in-itself", I think Nietzsche refers to people's existence in the nature of reality, which is in contrast with the way people impose their own constructs, such as causality, necessity, laws, and moral concepts, onto themselves.

*mythologically.* The "unfree will" is mythology; in real life, it is only a matter of *strong* and *weak* wills. It is almost always a symptom of what is lacking in himself when a thinker senses in every "causal connection" and "psychological necessity" something of constraint, need, compulsion to obey, pressure, and unfreedom; it is suspicious to have such feelings—the person betrays himself. (*BGE* 21)

There is something worth noting in this paragraph. One thing is that Nietzsche seems to change his attitude toward necessity. I argued that in *HH* Nietzsche holds the view that determinism rules out responsibility because we are not responsible for our nature. By contrast, he claims in *BGE* that "causal connections" and "necessity" are merely notions that we construct. We devised these items and acted upon them as if they existed in themselves. Another point is that an agent betrays herself to make an inference from "causal connection" and "psychological necessity" to unfreedom and obeying compulsion. If Nietzsche no longer believes in necessity and holds the inference from necessity to unfreedom as fictitious, then Nietzsche was no longer an incompatibilist at the time he wrote *BGE*.

So to sum up, Nietzsche's new form of argument can be stated as follows:

(1) An agent has free will in the superlative metaphysical sense and ultimate responsibility only if she is *causa sui*.

(2) No one can be *causa sui*. (For *causa sui* is a self-contradiction)

(3) Therefore, no one has free will in the superlative metaphysical sense and ultimate responsibility for her actions.

Moreover, Nietzsche draws a line between different kinds of free will and rejects the contrary of free will in the superlative metaphysical sense, namely "unfree will" as it also relies on the fictitious "causal connections" and "necessity". As I argued before, Nietzsche has already initiated this differentiation since he wrote *HH*, because he hoped to attain free will, while considering the free will available at the time to be immature and constrained by the chain of necessity. Taken together, we can see that Nietzsche was never, as Leiter thinks, the one arguing against free will, but looks forward to the final realisation of free will. Even though he was a determinist and source-incompatibilist before, he becomes sceptical of necessity and regards this transformation as a hope for free will. However, whether he believes in other kinds of free will is still unclear. It is worth noting that he proposes a distinction between strong and weak wills instead of free and unfree wills. This is an essential point that Nietzsche maintains in his later works.

### 3.1.3 *Twilight of the Idols* (1889)

There appear to be inconsistencies in Nietzsche's notion of free will and determinism in *Twilight of the idols*, which is considered a work of his late or mature period. On one hand, Nietzsche seems to echo the points he made in *HH*, holding the view of determinism. He claims that

> No one *gives* a human being his qualities: not God, not society, not his parents or ancestors, not *he himself* (—the nonsensical idea here last rejected was propounded, as 'intelligible freedom', by Kant, and perhaps also by Plato before him). *No one* is accountable for existing at all, or for being constituted as he is, or for living in the circumstances and surroundings in which he lives. The fatality of his nature cannot be distangled from the fatality of all that which has been and will be... One is necessary, one is a piece of fate, one belongs to the whole, one is in the whole... *But nothing exists apart from the whole!* (*TI* VI: 8)

In this passage, Nietzsche seems to maintain his view in *HH* that one is not responsible for his environment and existence so that one lacks responsibility. One is part of the whole universe, where everything that happens is determined by prior events and the laws of nature. One's actions cannot be traced back to one's control. Therefore, one is necessary and does not have a moral responsibility. The "freedom of the will" is just an illusion.

However, what Nietzsche emphasises here is the Kantian notion of "intelligible freedom", not all kinds of freedom. In other passages in *TI*, he distances himself from the ideas in *HH* and reaffirms those proposed in *BGE*. Nietzsche continues his rejection of *causa sui* as self-contradictory. He writes that "All supreme values are of the first rank, all the supreme concepts... all that cannot have become, *must* therefore be *causa sui*. But neither can these supreme concepts be incommensurate with one another, be incompatible with one another" (*TI* III: 4). Additionally, in section 33 of "Expeditions of an Untimely Man", Nietzsche writes that the individual is not "a 'link in the chain', something merely inherited from the past" (*TI* IX: 33). In section 38, he then proposes his own conception of freedom:

> For what is freedom? That one has the will to self-responsibility. That one preserves the distance which divides us. That one has become more indifferent to hardship, toil, privation, even to life. That one is ready to sacrifice men to one's cause, oneself not excepted. (*TI* IX 38)

This passage clearly shows that Nietzsche explicitly proposes an alternative notion of freedom. What he truly rejects is not freedom itself, but rather necessity and the 'modern concept of freedom' (*TI* IX: 41), which are invented to make people accountable and impose the rewards and punishments associated with their actions. For Nietzsche, modern freedom is the degeneration of instinct, but there exists another form of freedom that represents the ascending line of life. This view echoes Nietzsche's shift in perspective in *BGE* and explains his repudiation of "ultimate responsibility" and freedom of the will "in the superlative metaphysical sense". Therefore, we cannot directly conclude that there are inconsistencies in Nietzsche's conception of free will across his texts, as he introduces his own notion of freedom, which makes it possible for the will to be free. Furthermore, Nietzsche maintains his distinction between strong and weak wills. Weak will, as Nietzsche defines it, is the "incapacity to resist a stimulus" (*TI* VIII: 6). In contrast, strong will is "not to react immediately to a stimulus, but to have the restraining, stock-taking instincts in one's control… the essence of it is precisely not to 'will', the ability to defer decision" (Ibid.). This distinction can be correlated with the condition of "sovereign individual" (*GM* II: 2), which I will explore in the next chapter, to assess whether the ability to control oneself is any sort of condition for free will.

## 4. Conclusion

To summarise, Leiter presents two models to argue that Nietzsche does not endorse any form of free will. I argue that the epiphenomenal model fails because it lacks support both in Nietzsche's texts and in empirical science. The secondary-cause model, along with the challenges posed by incompatibilism, is likewise unconvincing. Upon examining three of Nietzsche's works, I found that he was initially a determinist and source-incompatibilist when writing *HH*, but later underwent a shift, rejecting not only determinism but also the inference from necessity to unfreedom. In fact, Nietzsche advocates for a form of free will distinct from the one constrained by necessity. Although he initially accepted necessity, he later realised that the notions of necessity, causality, and even free will, which is framed in terms of praise and blame, function primarily as instruments of accountability and control. Therefore, he rejects only the conception of free will that presupposes the self-contradictory notion of *causa sui*. What he is struggling for is to liberate himself from these illusory notions and to pick the ripest fruit, the "mature freedom of spirit which is equally self-mastery and discipline of the heart" (*HH* I: P: 4). Moreover, Nietzsche replaces the dichotomy of free and unfree will with the conceptions of strong and weak will. If proposing strong and weak will instead is a more adequate way to approach free will, it is essential to examine whether this reimagining is connected with free will in other senses. In the following chapters, I will go beyond the traditional framework of the free will and moral responsibility debate to argue that Nietzsche, despite being widely known as an immoralist, advocates for a specific form of moral responsibility. This form of responsibility, in turn, constitutes the core of his ideal conception of free will.

# Chapter II

# Strawson's Naturalistic Turn

## 1. Introduction

In the first chapter, I argued against Leiter's account and proposed that Nietzsche is not, as Leiter thinks, always an incompatibilist; rather, he underwent a transformation in his thought and may advocate a specific notion of free will that is fundamentally distinct from the traditional understanding, which is constrained by necessity. Nietzsche also introduces the idea of strong and weak wills to to support his normative ideal of free will—that is, a conception of free will that he praises and endorses. The question then arises: what is Nietzsche's ideal notion of free will, and how might it be delineated? It is challenging to directly define Nietzsche's conception of free will, as evidenced by the attempts of various Nietzsche scholars. For example, some scholars employ the standard model to explain how genuine actions are produced according to Nietzsche. Maudemarie Clark and David Dudrick (2009) argue that Nietzsche's drives stand in an "order of rank" (*BGE* 6) based on normative authority rather than mere strength, and that one's willing is constituted by the normative order of drives. However, this model is vulnerable to the threat of alienation, meaning that one may be alienated from one's own beliefs and desires, resulting in the primary reason offering no justification for an action. To address the problem of alienation—where one may be estranged from one's own beliefs and desires, Katsafanas (2011) suggests that Nietzsche regards power as the constitutive aim of an action, which is legitimately binding on us. As a result, the agent's actions are in alignment with their desires, directed towards the realisation of power, thereby mitigating the threat of alienation. By actively committing to the pursuit of power and continuously overcoming resistance, the agent integrates their drives into a unified agency. Nietzsche's conception of freedom thus comes from the agent's commitment to power. This theory also encounters difficulties, because the aim of power can be applied to all human activity, including individuals who are not able to achieve true freedom. For instance, both sovereign individuals and priests may pursue power, but only the former can achieve freedom (*GM* II: 2), while the latter may seek power but remain alienated from their true will. Therefore, Katsafana's approach fails to identify a subset of actions that are highly demanding and rarely achieved. These attempts to interpret Nietzsche's account of free will merely illustrate a broader issue. Attempts to describe Nietzsche's concept of free will highlight the cyclical nature of scholarly debate, where each new theory tries to resolve previous problems but inevitably exposes new ones, making the conceptual construction seem like an intellectual game of refinement.

To address this issue, I will propose a new interpretation of Nietzsche's free will and moral responsibility. As I have previously argued, Nietzsche repudiates the traditional conceptions of free will and moral responsibility. In the later sections of this dissertation, I will demonstrate that Nietzsche regards free will and moral responsibility as capacities acquired through practice rather than as metaphysical constructs. His replacement of free and unfree will with strong and weak wills is an intentional redefinition that underscores the practical essence of the conceptions he advocates. However, in what sense can free will be understood as a strong will? I argue that, as a practical conception, free will should no longer be regarded as a prerequisite for moral responsibility but rather as its product. Like the sovereign individual Nietzsche praises, he becomes "master of a free will" (*GM* II: 2) only by acquiring the strength to break free from the constraints of traditional morality and proudly assume responsibility on the basis of one's own "measure of value" (*GM* II:2). Therefore, in order to understand Nietzsche's practical notion of free will, I will first focus on moral responsibility and, building on that foundation, attempt to delineate the core meaning of Nietzsche's radically re-interpreted notion of free will.

Nietzsche shifted from an incompatibilist stance as explored in Chapter 1. Moreover, in the second essay of *On the Genealogy of Morals*, he also depicted the sovereign individual as the highest expression of "the long history of the origins of responsibility", because of his "proud awareness of the extraordinary privilege of responsibility" (*GM* II: 2), suggesting that he does not repudiate all forms of moral responsibility. What kind of moral responsibility does Nietzsche affirm? How does moral responsibility acquire a practical dimension, and how does it contribute to the development of free will through the dynamic process of lived experience? When Nietzsche leaves his positive notion unclear and most scholars fail to elucidate the relationship between the two kinds of responsibility, his notion of responsibility seems to be a rhetorical expression that falls from the sky. Yet, Nietzsche views the sovereign individual who exhibits good qualities as "the ripest fruit" of the morality of custom. an ideal that emerges out of the very moral practices Nietzsche elsewhere critiques, indicating that the moral responsibility he values is not an arbitrary invention, but rather has a real historical and psychological origin within the morality he rejects. Moreover, when this figure is associated with terms like freedom, autonomy, self-mastery, and the capacity to make and keep promises—concepts that can be linked to the traditional metaphysical notion of moral responsibility, Nietzsche does not define these terms, but simply attributes them to the figure he sees as a exemplar model. This makes all of his praise for this ideal figure, and his promotion of the moral responsibility he attributes to it, vulnerable to being dismissed as mere rhetoric. As Leiter claims, "it seems to me a mistake, however, to read this passage as articulating a kind of ideal of agency or selfhood; in context, I think it is far more plausible to understand the passage as being wholly ironic and mocking." (Leiter, 2011).

What is often overlooked in current debates is the process through which the sovereign individual comes into being. In other words, in trying to show how the *positive conception of responsibility* arises within *GM* II's "long history" of responsibility, which is overwhelmingly concerned with moral responsibility Nietzsche rejects, people always overlook the psychological genesis of the sovereign individual, so their strategies often fail to explain how Nietzsche's positive conception of responsibility, as exemplified by the sovereign individual, does not implicate the form of moral responsibility he rejects grounded in *causa sui*. The sovereign individual, while an individual figure, does not stand outside this "history"—he exemplifies the qualities of moral responsibility that emerge through historical and psychological processes. Since this ideal of responsibility is ineluctably shaped by the very practices that gave rise to conscience—by this "long history"—any proponent of such a strategy must also offer a plausible account of the sovereign individual's place within it, particularly within *GM* II's genealogy of guilt. Failing to clarify the sovereign individual's position within the historical development of morality raises further questions. If the praise for the responsibility that the sovereign individual embodies is merely rhetorical, then what is the exact nature of the responsibility Nietzsche actually affirms, and how does it differ from the metaphysical notions he rejects? In other words, how does the sovereign individual exemplify a non-metaphysical model of moral agency? Clarifying the psychological structure embodied by a responsible agent is key to understanding the meaning of these words that Nietzsche uses to praise his moral ideal, as well as how these conceptions evolved historically and psychologically and built up the connections across different historical stages, thereby illuminating the positive notion of responsibility throughout Nietzsche's work.

Nietzsche's own philosophy, in fact, harbours the resources to resolve this explanatory challenge. There is ample evidence that he does not conceive of responsibility in metaphysical terms, but consistently criticises such notions, regarding them as fabrications of Christian morality designed to induce belief in guilt and free will. I argue that Nietzsche's self-designation as a psychologist reflects a crucial shift in his thinking—from questioning the metaphysical impossibility of moral responsibility to a psychological investigation of conscience, bad conscience, and guilt. As he claims in in *Ecce Homo*, "That a psychologist without equal speaks from my writings—this is perhaps the first insight gained by a good reader" (*EH*, "Why I Write Such Good Books", 5). Correspondingly, in *GM* II, Nietzsche turns his focus increasingly towards the development of conscience, bad conscience, and guilt—phenomena that are directly and explicitly linked to "responsibility" when that term is understood not metaphysically, but as a practice of holding oneself and others to affectively charged expectations. Guilt, in particular, plays a central role for Nietzsche because it expresses an agent's sense of being responsible or accountable for failing to act on certain reasons or motives. This shift explains why Nietzsche moves his attention away from criticising moral responsibility at a metaphysical level and instead explores the underlying psychological dynamics behind moral responsibility within practices in *GM* II. The central role he assigns to guilt reveals why he believes it is still possible to be 'responsible' without satisfying libertarian conditions of freedom. For Nietzsche, to be morally responsible for an action does not depend on whether one could have metaphysically chosen to act on moral motives, but on our moral practices and corresponding psychological states.

However, while Nietzsche offers a rich and insightful psychological portrait of moral development in *GM* II, particularly in his 'psychologist' perspective on the origins and development of conscience, guilt, and responsibility, what he provides is a poetic and assertive psychological picture rather than a systematic or analytically rigorous theory. In other words, his descriptions do not constitute a clearly constructed theoretical system, but rather a non-defined mode of thought that embraces tension. Nietzsche deliberately avoids a systematic, argumentative style in favour of metaphor, simile, rhetoric, and contrast. For example, his description of the "sovereign individual" is full of passion and symbolism: "This emancipated individual, with the actual right to make promises, this master of a free will…" (GM II: 2). But he does not specify what 'freedom' and 'will' are, nor does he explain how such a will arises, nor does he distinguish between the possible meanings of these terms in different contexts. Yet he never explains what he means by 'freedom' or 'will', how such a will might come about, or how these terms might be understood differently in distinct contexts. Similarly, when tracing the 'history' and psychological development of responsibility, Nietzsche does not articulate a theoretical framework identifying which psychological states constitute necessary or sufficient conditions for responsibility. Nor does he explain the mechanisms by which such states might arise. Consequently, although Nietzsche's text is rich in resources for responding to the interpretive challenges raised above, these resources are often obscured by the ambiguities and tensions he deliberately creates in his rhetoric. Building on Nietzsche's rich but elusive psychological insights, I suggest that P. F. Strawson's influential account in *Freedom and Resentment* can provide a clarificatory restatement of Nietzsche's psychological insights, allowing us to better understand responsibility as a practice embedded in our emotional life.

To substantiate my claim, it is helpful to highlight the significant parallels between Nietzsche's approach to responsibility and P. F. Strawson's influential analysis in "Freedom and Resentment". Specifically, both thinkers move away from metaphysical accounts of responsibility and instead focus on the interpersonal practices and emotional attitudes that constitute our experience of holding ourselves and others responsible. While Nietzsche does not provide a systematic theory to respond to the explanatory challenge, Strawson's framework offers conceptual clarity and a vocabulary that can illuminate the emotional and psychological structures underlying Nietzsche's description of various stages of moral development. In this way, Nietzsche's ambiguous uses of the term "responsibility" and those ideals linked to this conception are unified by the underlying attitudes psychological states presuppose, thereby making a coherent response to the aforementioned challenge possible.

What does it mean, more precisely, to say that Strawson's approach provides a clarificatory restatement of Nietzsche's description? The key connection lies in their shared focus on the practice of holding responsible and the emotions involved in that practice. In his work "Freedom and Resentment", Strawson turns away from abstract theorising to the everyday practices of holding responsible, which is the idea known as the naturalistic turn. Specifically, Strawson "depends less on a conceptual analysis of 'freedom' and more on a descriptive psychology of human moral emotions" (Russell, 2017: 69). His primary concern is not to identify the necessary and sufficient conditions for moral agency. Instead, he examines the distinctive affective responses—resentment, indignation, guilt—that structure our moral experience and constitute our practices of holding responsible. To do so, Strawson reverses the order of explanation by taking practices of holding responsible as the condition for one's state of being responsible. In other words, the practices of holding responsible determine what it means to be morally responsible. According to Strawson, although something is indeed missing in the traditional debate, he does not fill this gap through a conceptual analysis of free will and moral responsibility; instead, he turns to our personal feelings and reactions when we hold others responsible. What is vital here is that we should not "over-intellectualise" the free will debate (Strawson, 2008: 35).

However, Strawson's account faces challenges, which pose the threat that Strawson falls back into the traditional moral responsibility/determinism debate. To accrue benefits from Strawson's approach within my own interpretion of Nietzsche's texts, I must outline potential avenues for addressing the challenges his account faces. Therefore, the primary focus of this chapter is expository, though it is not exclusively so. Firstly, I will provide an overview of the central idea in Strawson's "Freedom and Resentment", introducing his arguments against incompatibilism. Then, I will pay attention to his naturalistic turn and how others influenced by him resolve the main problem associated with this idea. I conclude this chapter by outlining two potential solutions to this problem: the language-game approach and the "borderline-style answer". In the next chapter, I will apply Strawson's approach to interpreting Nietzsche's texts, focusing on the second essay of *On the Genealogy of Morals*, so as to investigate how Nietzsche constructs his ideal notion of moral responsibility through a descriptive account of actual human moral practice akin to Strawson's method, rather than through mere conceptual analysis.

## 2. Freedom and Resentment

### 2.1 The Dichotomy between Compatibilism and Incompatibilism

The foundation of Strawson's idea lies in the dichotomy between two opposing views in the free will/moral responsibility debate. These are compatibilism and incompatibilism, which Strawson labels as optimism and pessimism, respectively. The optimist, or classical compatibilist, seeks to resolve this issue by grounding the justification of moral responsibility in the regulation of behaviour. For the optimist, freedom is nothing more than an agent's ability to act according to her will despite compulsion and coercion, such as accident, mistake, ignorance, innate incapacity or psychological disorder. These factors not only limit freedom but also mitigate the force of moral condemnation and punishment, as the effectiveness of regulating behaviour in desirable ways depends on their absence. Thus, freedom is regarded as the absence of certain factors. As long as an agent can be appropriately influenced by the incentives of rewards and punishments and act according to the determination of her will, she can be regulated in a socially desirable way and considered to possess the freedom required for moral responsibility. Therefore, in classical compatibilism, freedom is the ability to bring about actions despite constraints that would otherwise undermine moral condemnation. On this account, determinism poses no threat to freedom, as there appears to be no conflict between the regulation of behaviour and the truth of determinism.

On the other hand, the incompatibilist, or the pessimist, argues that the optimist bases moral responsibility on the wrong foundation, as mere rewards and punishment are insufficient to justify this concept. The optimist adopts a utilitarian understanding of moral responsibility, which indicates the efficacy of regulation. However, even animals and children, who are not regarded as fully responsible agents, can be influenced by rewards, punishment and condemnation, thereby exhibiting desirable behaviour. In this sense, moral responsibility is nothing more than an instrument or tool to regulate social behaviour. However, for an agent to be morally responsible and subject to moral praise or condemnation, she has to really deserve it. According to the pessimist, what the optimist is missing is the condition of an agents's deserving praise and blame, and this condition of moral desert goes beyond the freedom granted by classical compatibilism. That is, the agent must be the ultimate source of his actions, which is incompatible with the truth of determinism. Those who believe this condition can be met are known as libertarians, whereas hard determinists argue that determinism renders freedom and moral responsibility impossible.

According to Strawson, in order to provide a coherent account of freedom and moral responsibility that diverges from the optimist's perspective, libertarians tend to take anti-naturalistic position, asserting that agents must possess "contra causal freedom" (Strawson, 2008: 35)—a type of freedom that is immune to determinism. This view requires a "metaphysical" solution, as it posits that agents must be the ultimate originators of their actions, independent of causal antecedents. However, such a demand appears to be unattainable. What Strawson aims to do is to reconcile the opposition between the optimist and the pessimist and find a middle ground between them. Rather than adopting the pessimist's approach, which frames moral responsibility as a metaphysical construct, he draws our attention to moral practices and argues that moral responsibility is manifested in attitudes such as anger and resentment, which are inevitable human emotions. Therefore, in "Freedom and Resentment", Strawson focuses on our commonplaces and is concerned with the link between reactive emotions, which I will explore in detail later, and the practice of holding others responsible. "The object of these commonplaces is to try to keep before our minds something it is easy to forget when we are engaged in philosophy, especially in our cool, contemporary style, viz. what it is actually like to be involved in ordinary interpersonal relationships, ranging from the most intimate to the most casual" (Strawson, 2008: 23). The optimist's approach is right in emphasising the efficacy of practices that regulate behaviour and express our moral attitudes, but it overlooks the emotional dimensions inherently embedded in human practices, treating them merely as tools for social regulation rather than as genuine expressions of our moral nature (Strawson, 2008: 36). Libertarians see the lacuna in the optimist's account, but try to find an obscure metaphysical ground for those emotions. Hence, Strawson's strategy to address these deficiencies is to modify the optimist's framework "without recourse to the obscure and panicky metaphysics of libertarianism" (Ibid.) by shifting the focus away from the conceptual analysis of "freedom" and "responsibility" and focus on "that complicated web of attitudes and feelings which form moral life as we know it" (Strawson, 2008: 34). What is essential is "the very great importance that we attach to the attitudes and intentions towards us of other human beings, and the great extent to which our personal feelings and reactions depend upon, or involve, our beliefs about these attitudes and intentions" (Strawson, 2008: 22).

### 2.2 Reactive Attitudes and Quality of Will

As mentioned above, Strawson's strategy is to persuade the pessimist to reconsider by fundamentally reinterpreting the optimist's framework. In doing so, he does not seek to establish sufficient or necessary conditions of moral responsibility. Rather, he takes a "naturalistic turn" and shifts the focus towards the practices of holding responsible. Specifically, Strawson "depends less on a conceptual analysis of 'freedom' and more on a descriptive psychology of human moral emotions" (Russell, 2017: 69). His primary concern is the relationship between moral responsibility and specific human emotions and affective responses, rather than the necessary and sufficient conditions of moral agency.

Central to Strawson's account is his concept of reactive attitudes, such as resentment, gratitude, forgiveness, guilt, contempt, and love. For Strawson, these emotions are inescapable, as they are "something as deeply rooted in our natures as our existence as social beings." (Strawson, 2005: 34). The practices through which we hold others responsible, including praise, blame, as well as the role of excuses and exceptions, which I will elaborate on below, can be understood through these inescapable emotions, which constitute a fundamental aspect of human life.

Then what exactly is reactive attitude? Strawson writes that reactive attitudes "are essentially reactions to the quality of others' wills towards us, as manifested in their behaviour: to their good or ill will or indifference or lack of concern" (Strawson, 2008: 28). Strawson first considers instances of resentment, namely, cases in which a person is offended or injured by another without goodwill. Under these conditions, one would naturally feel resentment if the offender does not show any kind of regret or awareness of the wrongdoing. Similarly, if one is helped by another person simply out of goodwill, rather than accident, the recipient will naturally feel the emotion of gratitude. Such situations imply the essential role reactive attitudes play in shaping our interpersonal relationships, because "we demand some degree of goodwill or regard on the part of those who stand in these relationships to us" (Strawson, 2008: 23) and this grounds the justification for moral responsibility. Given our fundamental concern with how others treat us, we naturally question "whether the actions of other people—and particularly of *some* other people— reflect attitudes towards us of goodwill, affection, or esteem on the one hand or contempt, indifference, or malevolence on the other" (Strawson, 2008: 22). Put differently, reactive attitudes express "a certain sort of demand for inter-personal regard" (Strawson, 2008: 30).

It is important to note that although we expect people to act in certain ways, our concern is not limited to their behaviour alone, but rather their intentions and motivations that guide their behaviour. In other words, we care about someone's "quality of will": whether she demonstrates concern for others, as opposed to ignorance or disregard. According to McKenna, a person's quality of will is shaped by "the regard or concern one has toward others (or oneself), and toward the relevance of moral considerations, as manifested in one's conduct" (McKenna, 2012: 59). Our reactive attitudes toward others' behaviour actually take their quality of will as the target. Even in situations where we are initially upset by someone's actions, our feelings tend to diminish or disappear once we realise that they did not do so with bad intentions. As will be explained in the next section, if someone accidentally drops something that causes us harm, we might feel unlucky or upset, but our anger towards her is likely to lessen or vanish once we understand it was an accident caused by the slippery floor. Therefore, it is the intentions behind the behaviour, rather than the behaviour itself, that we truly care about.

In conclusion, Strawson holds that reactive attitudes are central to our various connections with others. They reflect the "importance we attach to the attitudes and intentions towards us of those who stand in these relationships to us" (Strawson, 2008: 23). When someone fails to live up to the expectation, we are prone to emotions such as resentment, anger and moral condemnation. When someone acts out of kindness, we are prone to experience gratitude and appreciation. These reactions, in turn, lead to specific behaviours and dispositions, shaping how we interact with others in social contexts

### 2.3 Excuse and Exemption

I have investigated the nature of reactive attitudes. Although it is difficult to give a precise definition of reactive attitudes, they are practical attitudes describable psychologically or phenomenologically but not fully captured by concepts, it seems that the core of this notion implies that we are inclined to hold people responsible when experiencing an episode of emotions, which are directed towards the responsible agent. However, under certain conditions, a person is disposed to modify or even suspend his feeling of resentment. These situations are common in our daily lives since we are not always justified in holding others responsible based on our emotional reactions to their behaviour. For example, if my friend reveals my secret to others, I would hold her responsible for not taking sufficient care in maintaining the confidentiality of the information, and I would feel resentment towards her as a result. However, if she explains that the secret was revealed because her phone was stolen, despite her taking great care, I would be able to alleviate my resentment and recognise that she did not do so on purpose, and therefore, she should not be responsible for the situation.

Strawson distinguishes two types of conditions, in which our reactive attitudes are modified or mollified because of new information about others' quality of will. The first group of pleas include expressions such as "He didn't mean to", "He hadn't realised", "He didn't know", and also "He couldn't help it", "He was pushed", "He had to do it", "It was the only way", "They left him no alternative" etc (Strawson, 2008: 23). What these pleas share is that they do not suggest the complete suspension of our ordinary reactive attitudes, but instead regard one particular attitude as inappropriate. To be more specific, in these kinds of excusing conditions, what influences our reactive attitudes is not the moral status of an agent because the actual quality of her will is not what we initially assumed she acted with, which renders our reactive attitudes towards her actions inappropriate. In specific situations, an agent's actions may lack any kind of ill will. They are merely accidental, in that she was unaware of the consequences her actions caused, or she acted out of reasons that outweighed her reluctance, so "they do not invite us to see the agent as other than a fully responsible agent", but "invite us to see the injury as one for which he was not fully, or at all, responsible" (Ibid). She remains a fully responsible agent and an appropriate target for the demand for goodwill. For instance, if I learn that my friend revealed my secret accidentally, rather than intentionally, I would modify my reactive attitude, as I now have a better understanding of her quality of will and recognise that it was not malicious.

The second category of pleas is commonly referred to as "exempting conditions". For Strawson, sometimes people are exempt from the demands our reactive attitudes express. He classifies these situations into two distinct sub-groups. The first one contains expressions such as "He wasn't himself", "He has been under very great strain recently", "He was acting under post-hypnotic suggestion" (Strawson, 2008: 24). We do not typically experience our usual reactive attitudes towards individuals in exceptional circumstances, since they are not the person they should have been when they acted so. The second sub-group assumes the situations in which an agent is abnormal or morally undeveloped. It is not uncommon for us to make exceptions for an individual's wrongdoings, as seen in statements like "He's only a child", "He's a hopeless schizophrenic", "His mind has been systematically perverted" (Ibid.). When we see a person in such a light, our reactive attitudes will be profoundly modified because we do not hold them to be in a condition allowing them to be morally responsible for their behaviour. Underlying these sub-groups of exempting conditions is that the full suspension of reactive attitudes results not from lack of ill will, but from the impossibility of agency given their state, be that temporary or longterm. She is not an appropriate target for interpersonal relationships and the demand for goodwill when she is under abnormal strains, neurotic or morally undeveloped, all of which invite us to view her in a different light from the light "in which we should normally view one who has acted as she has acted" (Ibid.). In this situation, we should no longer adopt "the attitude of involvement or participation in a human relationship" (Ibid.), but should take an objective attitude, in which we no longer perceive the agent as an ordinary human being but rather as an object of social policy, someone who is a subject that can be "managed or handled or cured or trained" in the natural world. In essence, we observe the person from a different standpoint, because she is hindered in forming regular interpersonal connections. Taking an objective attitude towards those who are not morally competent is, therefore, to regard them as exempt from moral demands and, to some extent, to set them apart from the moral community. They can still have emotions such as repulsion, fear, pity and some kinds of love, but they are not recipients of emotions that belong to communication with others in inter-personal relationships. In other words, they are not, "to that extent, seen as a morally responsible agent, as a term of moral relationships, as a member of the moral community" (Strawson, 2008: 30).

To sum up, excusing and exempting conditions shed light on the possibility that sometimes we may occasionally misattribute responsibility based on our reactions to others' behaviour. Alternatively, one could argue that reactive attitudes are not appropriate in these situations. Regarding excusing conditions, what makes reactive attitudes inappropriate is that excuses provide extra information about the quality of will of the one who is held responsible. When I blamed my friend for revealing my secret, I assumed she did not value my trust or the importance of keeping my secret, nor did she respect his promise to me. This mistakenly indicates her quality of will because later, it turns out that she does care about my secret, but can do nothing after her mobile was stolen. The quality of an individual's will plays a crucial role in regulating moral attribution. In exempting conditions, the exempted individual lacks the capacity to exhibit their quality of will due to psychological or moral limitations. If they fail to grasp the meaning of moral demands, they cannot exhibit any quality of will, which, in turn, diminishes our reactive attitudes due to the absence of a morally accountable subject. What connects excusing and exempting conditions sis their intrinsic link to the concept of quality of will.

### 2.4 Strawson's Argument Against the Pessimist

Recall the dichotomy between the optimist and the pessimist. The optimist holds that moral responsibility has its justification in regulations of behaviour. The beneficial consequences of regulating behaviour are sufficient to to renderthe truth of determinism irrelevant. The pessimist correctly argues that the optimist overlooks something vital and seeks to fill in the lacuna, but they do so by engaging in metaphysical investigation and claim that moral responsibility is justified only if we possess certain capacities incompatible with the truth of determinism. According to Strawson, there is no need for the pessimist to go beyond the facts we know because "the vital thing can be restored by attending to that complicated web of attitudes and feelings which form an essential part of moral life as we know it" (Strawson, 2008: 34). For him, reactive attitudes are given to us as human beings, and we are thus inescapably committed to feeling them even in a deterministic world. Strawson presents several arguments against the pessimist to demonstrate the significance of reactive attitudes and moral responsibility in our lives. I will discuss three of them in turn.

### 2.4.1 The Naturalistic Argument

The Naturalistic Argument, in my view, is central to Strawson's critique of the pessimist. It highlights the nature of reactive attitudes and asserts that we are inescapably committed to them, with no means of abandoning them. Strawson supports this argument in two ways. On the one hand, reactive attitudes are "too thoroughgoing and deeply rooted for us to take seriously the thought that a general theoretical conviction might so change our world"(Strawson, 2008: 26). They are deeply ingrained emotions that we naturally experience in everyday life, forming an integral part "of the general framework of human life" (Strawson, 2008: 28), so that the general framework of reactive attitudes "neither calls for, nor permits, an external 'rational' justification" (Strawson, 2008: 35). Just as it is psychologically impossible for us to cease to feel happy and sad, so too is it impossible to entirely suspend reactive attitudes because they are in our nature. On the other hand, the suspension of reactive attitudes would requires "adopting a thoroughgoing objectivity of attitude to others" (Strawson, 2008: 27), which is beyond our human capability. As Strawson writes: "A sustained objectivity of inter-personal attitude, and the human isolation which that would entail, does not seem to be something of which human beings would be capable, even if some general truth were a theoretical ground for it" (Strawson, 2008: 26). This explains why taking an objective attitude towards those in abnormal circumstances does not imply that we should treat them as mere objects. We are willing to accept them in interpersonal relationships, but their rare abnormality sets them apart from such involvement. Moreover, while we may invoke the objective attitude as a resource, we cannot "do this for long, or altogether" in normal cases (Strawson, 2008: 25). Thus, determinism, even if it is true, is idle or irrelevant to our concept of moral responsibility, because it is impossible for us to fully adopt the objective attitude and abandon reactive attitudes, which are fundamental expressions of our moral nature.

One of the most significant responses to Strawson's naturalistic argument is the one put forward by Paul Russell. Russell (1992) argues that Strawson's argument for the inescapability of reactive attitudes fails because he does not distinguish between type and token-naturalism. Type-naturalism posits that our liability and disposition to reactive attitudes are intrinsic to human nature and require no general justification of any sort. There is no way for us to disengage from these types of natural inclinations or emotions. Russell develops type-naturalism further by drawing an analogy between fear and reactive attitudes. Our susceptibility to the emotion of fear, he argues, is a basic fact of human nature. Just as there is no need to justify the fact that the humans are born with a heart and two kidneys, so too is there no need to justify our susceptibility to fear. The analogy between reactive attitudes and fear supports the view that reactive attitudes are natural facts given to us, which require no external justification.

However, what if the pessimist is a token-pessimist rather than a type-pessimist? According to token-pessimism, although we may be liable to fear, "we are nevertheless capable of altogether ceasing to feel or experience fear if and when we judge that, given our circumstances, this emotion is never justified" (Russell, 1992: 294). In other words, there is no conflict between type-naturalism and token-pessimism, making type-naturalism unable to serve as a response to token-pessimism, because token-pessimism does not challenge our inclination to feel reactive attitudes, as type-naturalism implies, but argues that, in certain contexts, every token of reactive attitudes may be unjustified. We are equipped with the mechanism to feel reactive attitudes, but we are not justified in utilising this mechanism in these situations. For Russell, a deterministic world provides such a context, in which we are never justified in entertaining episodes of reactive attitudes because all of us are determined.

To counter token-pessimism, Strawson would need to adopt token-naturalism. Nevetheless, token-naturalism is a stronger line than type-naturalism, asserting that no reason of any sort can make us completely cease feeling afraid. It is evidently false that we are incapable of ceasing to entertain episodes of reactive attitudes "in circumstances where we believe that it is never appropriate or called for" (Russell, 1992: 295). For instance, in cases of excuses and exemptions, we may modify or even mollify our reactive attitudes towards those who are inappropriate targets of moral attributions. Thus, token-naturalism denies our moral practices any role in determining when moral attributions are appropriate in certain contexts, which is far from truth. As a result, Strawson faces a dilemma. He must either embrace type-naturalism, which cannot effectively discredits token-pessimism, or adopt token-naturalism, which denies an evident feature of moral practice. To discredit pessimism using Strawson's naturalistic argument seems implausible.

### 2.4.2 The Rationalistic Argument

The rationalistic argument aims to show that the truth of determinism would not lead to situations in which everyone is excused all the time. Recall that Strawson examines the conditions under which we would modify or mollify reactive attitudes and outlines two types of pleas. The first kind, referred to as the excusing conditions, suggests that an agent acted without the presence of ill will. Her behaviour may result from a mistake, an accident, or coercion. But what if determinism is true? Does the truth of determinism generalise the excusing conditions? Strawson clearly gives a negative response because he believes it would never "follow from the truth of determinism that anyone who caused an injury *either* was quite simply ignorant of causing it *or* had acceptably overriding reasons for acquiescing reluctantly in causing it *or* …." (Strawson, 2008: 26). Determinism is irrelevant to concerns about moral attributions because living in a deterministic world does not imply that all actions are due to negligence or lack of ill will. In other words, determinism does not make excusing conditions a universal scenario; therefore, it does not require us to always modify or mollify reactive attitudes, rendering moral attribution impossible.

The second type of pleas are exempting conditions. An agent is exempt from moral attributions because she is regarded as an inappropriate target for demanding goodwill. Such persons are either psychologically abnormal or children who are morally undeveloped, and thus, they will not be viewed as a member of the moral community. It is preferable to adopt the objective attitude towards them and treat them as individuals to be trained or cured. If it is psychological abnormality and undeveloped moral sense that exempt one from moral demands, would the truth of determinism generalise these exempting conditions? Strawson attempts to show that determinism does not imply that everyone is always abnormal or morally undeveloped. According to Strawson, the possibility of universal abnormality is self-contradictory because it would become normal once applied to all persons. Even if the the possibility is not self-contradictory, we still cannot take up the objective stance towards everyone at all time, regardless of the theoretical view we hold.

According to Russell (1992), the rationalistic argument presented by Strawson is still inadequate. He focuses on Strawson's exempting conditions and argues that Strawson's argument conflates "abnormality" and "incapacity". For Strawson, abnormality can in no way become a universal condition. However, what underlies the two exempting conditions is actually people's incapacity to express their reactive attitudes when they are psychologically abnormal or morally undeveloped. If we replace Strawson's terminology "abnormality" with "incapacity", we will find that "it is not inconceivable or self-contradictory to suggest that there could be a world, or things might develop, such that everyone is or becomes incapacitated" (Russell, 1992: 299). It is possible that a severe disease would make all of us incapacitated. If this could be the case, then the pessimist should be understood as claiming that determinism would not make us morally abnormal, but morally incapable, which would make us inappropriate targets of moral attributions. There is nothing self-contradictory in the thesis that incapacity could be universal.

Moreover, Strawson's account of the excusing conditions is also under threat. Recall Strawson's claim that the truth of determinism in no way leads to the consequence "that anyone who caused an injury *either* was quite simply ignorant of causing it *or* had acceptably overriding reasons for acquiescing reluctantly in causing it *or* …." (Strawson, 2008: 26). Strawson does not complete the list, but when examining some examples he proposes, such as "He couldn't help it", "He was pushed", "He had to do it", "It was the only way", "They left him no alternative", we will find that their common foundation for excusing an agent depends on whether one could have done otherwise. Then we can reconstruct Strawson's argument as follows: it would not follow from the truth of determinism that anyone who caused an injury either was quite simply ignorant of causing it, had acceptably overriding reasons for acquiescing reluctantly in causing it, or could not have done otherwise. However, this is what the pessimist insists about the truth of determinism. The pessimist argues that determinism is incompatible with the ability to have done otherwise, but Strawson contends that determinism would not have such a consequence because the incapacity could not be universalised. Strawson, in this sense, is begging the question.

Therefore, it appears that neither the excusing nor the exempting conditions are effective in responding to the pessimist. It is open for the pessimist to claim that determinism does indeed generalise the excusing conditions Strawson outlines, such as the incapacity to have done otherwise. They may also contend that the truth of determinism would pose a threat to our capacities as moral agents, thereby rendering us all incapable of bearing responsibility.

### 2.4.3 The Normative Argument

Strawson acknowledges that his naturalistic and rationalistic argument may not sufficiently address the pessimist's objections, but he insists that even if we have to abandon reactive attitudes in practice, which would be impossible for him, and adopt a thoroughgoing objective attitude towards others, "our doing so is not the consequence of a theoretical conviction which might be expressed as 'Determinism in this case', but is a consequence of" different reasons. More specifically, our decisions about what to do should only be made "in the light of an assessment of the gains and losses to human life, its enrichment or impoverishment; and the truth or falsity of a general thesis of determinism would not bear on the rationality of *this* choice" (Strawson, 2008: 27). Since being exposed to reactive attitudes builds up various interpersonal relationships and makes us members of a moral community, the abandonment of reactive attitudes would result in the loss of all these relationships, which is too high of a price to pay. In this regard, even if determinism is true, we should not give up our moral practices because doing so would be a great loss to humans.

From the above, it appears that for Strawson, participant reactive attitudes are necessary for meaningful relationships between humans. We cannot view ourselves as engaging in meaningful relationships with other moral agents unless we expect them to demonstrate some goodwill towards us. As the demand for goodwill is expressed through reactive attitudes, it follows that genuine meaningful relationships require reactive attitudes as necessary components. If we abandon reactive attitudes, we also forfeit our demand for goodwill towards us, thereby undermining all meaningful relationships with others. In this sense, detaching from reactive attitudes tends to foster objective attitudes.

However, Pereboom (2016) argues against the participant/objective attitudes dichotomy by contending that the suspension of participant reactive attitudes does not necessarily result in the objective attitude, because we can still sustain interpersonal moral relationships with others. For Pereboom, there are analogues that can do the same job as reactive attitudes, without presupposing that their targets are morally responsible agents, yet still capable of conveying relevant information in relationships. Pereboom illustrates this argument by giving several examples. One example is the attitude of forgiveness, which presupposes that the person being forgiven deserves blame in the first place. Determinism would, therefore, clearly undermine forgiveness because in a deterministic world, no one would deserve blame. However, Pereboom argues that "there are certain features of forgiveness that are not threatened by this view, and these features can adequately take the place this attitude usually has in relationships" (Pereboom, 2016: 229). For example, if someone has wronged you multiple times but apologises to you with recognition of their wrongdoing, the typical response in maintaining the relationship is "to overlook deserved blame or punishment" (Ibid.), which would be jeopardised by the conviction of determinism, as determinism undermines desert. However, it may also be possible for you to "simply choose to disregard the wrong as a reason to alter the character of your relationship", which presupposes the relinquishment of deserved blame or punishment, and is therefore compatible with determinism (Ibid.). In this case, the only component of forgiveness that is given up is the belief in deserved blame and punishment, which means that the willingness to overlook them is no longer required to sustain good relationships. In other words, an analogue can take the place of the role of forgiveness in interpersonal relationships.

What Pereboom is suggesting is that while reactive attitudes are essential for interpersonal relationships, they are not the only ones, and thus, the abandonment of reactive attitudes does not directly lead to the objective attitude. Not only is forgiveness like this, but reactive attitudes such as gratitude, which presupposes moral responsibility for praiseworthy actions, can also be replaced by analogous feelings like joy, which do not require consideration of moral responsibility, while still maintaining interpersonal relationships. If this is the case, then the normative argument seems to lose its force, because the loss to human life is not that significant after all.

## 3. The Naturalistic Turn and the Nature of Moral Responsibility

Strawson's response to pessimism remains controversial. Although his three arguments face their own issues and challenges, this does not necessarily mean that his response is a failure. For example, considering Pereboom's criticism of the normative argument, it is conceivable that, beyond the limited reactive attitudes listed by Strawson, additional emotions may also play a crucial role in sustaining interpersonal relationships. If so, substituting reactive attitudes with analogous emotions may not be as feasible as Pereboom suggests. Thus, there remains significant room for debate on whether Strawson successfully address the problem of pessimism. However, regardless of the specific details, Strawson's willingness to engage with pessimism reflects his most essential contribution to the problem of free will: his naturalistic turn, which shifts attention away from the metaphysical debate and toward human natural emotions. Before Strawson, much of the debate concerning free will, especially within analytical philosophy, is tightly in the grip of conceptual analysis. Resolving the issue was thought to depend on the correct analysis of these concepts, such as "free will", "moral responsibility", and "causal relation", under the assumption that the appropriateness of holding someone responsible requires establishing whether they are truly responsible. However, these analyses themselves do not see any hope of an end, because both optimists and the pessimists, according to Strawson, seek to "over-intellectualise the facts" (Strawson, 2008: 35). They continuously search for the conditions for free will and moral responsibility, and when confronted with objections, they attempt to identify conceptual flaws in their opponents' arguments while refining their own conceptual frameworks in response. This process results in an ongoing loop of conceptual revision without resolution. Strawson tries to fill in what is missing in the optimist's account without falling into the metaphysical debate by reversing the order of explaining the nature of moral responsibility. His naturalistic turn from metaphysical conceptions to human psychology represents a fundamental departure from the traditional philosophical methodology and grounds the justification of moral responsibility in our everyday practices of holding others responsible. In this way, Strawson does not explain moral responsibility in terms of metaphysical conditions one has to fulfil and blocks the pessimist's challenge that a condition for being responsible is the ability to do otherwise.

The contrast between Strawson's naturalistic turn and the traditional investigation into moral responsibility raises the question concerning the relationship between being responsible and holding responsible. To be more specific, is being responsible a property independent of our moral practices, or do our moral practices confer the property of being responsible? Strawson faces a challenge when he arrives at this crossroad. Recall that Strawson provides examples of excusing and exempting conditions to illustrate that, at times, an agent is not an appropriate target of moral attributions. Under excusing conditions, individuals commit wrongdoing without ill will. Under exempting conditions, they are psychologically impaired or morally undeveloped and therefore incapable of expressing any intentions through their behaviour. Both cases suggest that, at times, moral demands cannot be placed on individuals via reactive attitudes. This raises a critical issue. Is it the case that our reactive attitudes would wait for the decision of reason to determine whether an individual could be an appropriate target of moral blame, and then choose to make an outburst? Specifically, would our moral attitudes always be right in identifying the appropriate target? If Strawson is correct in asserting that holding responsible can account for being responsible, then if we were to hold very young children responsible and found ourselves unable to modify our emotions, would they truly be responsible?

It is conceivable to say that the mere practice of holding responsible is not sufficient for being responsible or not. There are more properties independent of moral practices that can render reactive attitudes inappropriate. However, if this is the case, then Strawson must either revert to the traditional debate on moral responsibility and have to account for the additional properties that determine an agent's moral responsibility, or he has to accept that young children can justifiably be subjects of moral blame.

Nomy Arpaly makes a similar critique, arguing that "contrary to those who think that blameworthiness and praiseworthiness of the agent for her action are epiphenomena of reactive attitudes, it is the blameworthiness or praiseworthiness of the agent for her action that makes her an 'appropriate' object of the emotions, and not the other way around. If we simply follow the moral emotions in order to figure out who is blameworthy and who is not, we are bound to be misguided" (Arpaly, 2006: 28). Emotions are unstable and subject to various external factors; thus, to avoid counterintuitive conclusions, an independent account of moral capacities may be necessary in determining responsibility.

This further reinforces the issue of the dichotomy between being and holding responsible. Fischer and Ravizza (1993) observe a distinction between holding responsible and being responsible and claim that the correlation between reactive attitudes and their legitimate objects may not be as strong as commonly assumed. It is conceivable that an entire society has its reactive attitudes switched on and off in the wrong way and at the wrong time. For example, under Nazism, individuals revelled in cruelty (e.g., military victories and the successful execution of genocide) and stirred up patriotic fervour for illusory aspirations (e.g., the revival of the glory of the Austro-Hungarian Empire). In retrospect, their reactive attitudes and emotions were set apart from their appropriate targets, leading to a state of moral chaos throughout society. In other words, moral responsibility appears to encompass more than simply being targeted by reactive attitudes. A deep problem related to this is that what makes an agent an appropriate target of reactive attitudes may also hinge on their broader moral capacities. Only with a well-defined interpretation of this moral capacity—for example, the capacity to participate in the moral community as a fully developed adult—can Strawson fill in the gap between reactive emotions and moral responsibility. This presents a challenge for Strawson because the pessimists may contend that the relevant capacities for a responsible agent is the ability to act otherwise, which appears to be "determinist contradictory" (Strawson, 2008: 35). Should Strawson reject the pessimists' interpretation of relevant capacities and engage in the debate by looking for other capacities compatible with determinism, he risks being drawn back into the endless conceptual analysis he initially sought to avoid.

Therefore, while Strawson's naturalistic turn is undeniably ambitious, it ultimately remains incomplete. If holding responsible determines the fact of being responsible, then holding young children responsible would, paradoxically, render them genuinely responsible. However, should Strawson attempt to identify moral capacities independent of moral practices, he risks reverting to conceptual analysis, which implies the traditional order of explanation and ultimately undermines the viability of his naturalistic turn. This leaves room for pessimists to argue that being responsible entails the ability to do otherwise. To respond to this challenge, Strawson has to formulate conditions compatible with determinism, which would draw him back into the traditional free will/ moral responsibility debate he originally sought to circumvent.

### 3.1 Strawsonians and Possible Answers

As argued above, the problem of appropriateness is central to the Strawsonian account of moral responsibility. By reversing the explanatory order of moral attributions, Strawson aims to demonstrate that conceptual analyses of free will and the truth of determinism are redundant, giving priority to the practice of holding responsible. However, as mentioned above, doing so gives rise to the problem of appropriateness, which requires Strawson to specify the moral capacities and conditions necessary for someone to become a target of moral reactions. If we attempt to identify these capacities, we leave room for the pessimist to argue that the truth of determinism undermines the validity of these conditions, thereby dragging us back into the trap of endless conceptual analysis. It seems that the only viable solution is to give up the explanatory priority of the practice of holding responsible and argue instead that the pessimist's conditions of being responsible, which are incompatible with determinism, are misleading. Thus, the challenge for any Strawsonian theory is to clarify the nature of appropriateness and avoid slipping back into the traditional debate. In other words, it is essential for Strawsonians to defend the position that we are inevitably committed to moral practices, a defence that can render the pessimist's concerns irrelevant, while also addressing what it means to be an appropriate target of reactive attitudes.

Some scholars have tried to extend Strawson's outline of moral responsibility, offering valuable attempts to address the problems I have identified above. These scholars shift our focus to the nature of moral responsibility and hold that Strawson's account contains the resources necessary to tackle these challenges, including the issue of appropriateness.

### 3.1.1 R. Jay. Wallace's Moral Reason

Wallace (1994) develops a theory of moral responsibility inspired by Strawson's naturalistic turn and examines the conditions of the appropriateness of holding someone responsible. But what does appropriateness mean according to Wallace? Wallace links reactive attitudes with the expectations people adhere to moral obligations and argues that moral justification provides reasons for complying with the obligation to meet expectations. For Wallace, holding individuals morally responsible for fulfilling moral obligations entails holding them accountable to these expectations, which is closely connected with reactive attitudes. We tend to experience negative emotions if our obligations or expectations have been breached. In this way, the pre-condition for being a morally responsible agent is the capacity required for grasping and being guided by moral reasons. Appropriateness, therefore, should be understood in terms of whether an expectation has been breached and whether the agent has the capacity to comprehend and apply moral reasons. To be considered a moral agent, one must possess the competence to understand and apply moral reasons, thereby being able to recognise moral demands tied to others' expectations and, in turn, be an appropriate target of moral attributions. Moreover, this competence, which is a necessary condition for being responsible, is not in tension with the conviction of determinism because, even in a deterministic world where all actions are predetermined, individuals can still possess moral understanding. If a deterministic world were to eliminate moral understanding altogether, this would imply the nonexistence of morality itself—a claim that is clearly at odds with reality.

Wallace's account offers a solution to the dichotomy between being and holding responsible by incorporating a cognitive component into our moral practices. Even if an entire society has its reactive attitudes switched on and off in the wrong way and at the wrong time, the stance of holding people to expectations still exists, with the appropriateness of holding someone responsible grounded in the possession of the normative competence required to grasp and apply moral norms that govern the expectations for goodwill. However, two issues arise in Wallace's account. First, it seems that for Wallace, the mere possession of moral capacity is sufficient to ground the justification of moral responsibility and address the pessimist's concern. However, it would be strange to hold someone responsible who has no control over the way she excises this capacity. Although the capacity required to comprehend moral norms may be compatible with determinism, it leaves room for the pessimist to argue that there is a tension between determinism and the ability to exercise this capacity. It is conceivable that, in a scenario where individuals are determined by external factors—such as in a particular kind of deterministic world—people may be unable to exercise the capacities they possess due to external constraints. This results in a more serious problem: when Wallace begins identifying the capacities for being morally responsible, he attempts to show that these capacities are compatible with determinism, which draws him back into the traditional debate Strawson aims to avoid. To be more specific, when Wallace presents his idea, he begins conceptually analysing whether a potential condition can be compatible with determinism, much like the pessimists do, rather than outlining the possibilities of free will and moral responsibility through the lens of moral practice in everyday life.

### 3.1.2 Gary Watson's Moral Address

Watson (2004) examines Strawson's claim that reactive attitudes essentially involve regarding the other as "a morally responsible agent, as a term of moral relationships, as a member of the moral community" (Strawson, 2008: 30). He introduces the idea of moral address and argues that holding someone responsible necessitates the validity of moral address, which in turn requires the agent to be regarded as a potential interlocutor within the community. For Watson, reactive attitudes represent forms of communication or address directed at a responsible agent, signalling whether she has done something right or wrong. So, what constitutes the content of reactive attitudes? Watson differentiates between basic concern and basic demand. As human beings, our basic concern lies in caring about how others treat and regard us and each other. The basic demand stems from this concern and expresses the hope of being treated with regard and goodwill. The basic concern and demand also structure our social relationships with others, defining what it means to be in interpersonal relationships and what is expected of us as moral agents within a moral community. Since our basic concern and demand are intrinsic to our nature and cannot be rationally justified, the framework they establish, within which the appropriateness of moral attributions makes sense, likewise cannot and does not need to be justified. In this way, the pessimist's challenge becomes irrelevant because whether one is morally responsible is not based on reasons but is given to us by nature; there is no reason to abandon moral practices even if determinism is true. Watson's interpretation of the communicative nature of reactive attitudes and his distinction between basic concern and demand suggests our nature constrains or determines who can be the appropriate target of moral attributions. As human beings, we are inescapably committed to the two fundamental properties of humanity, which structure the competence required for engaging in interpersonal relationships and, in turn, structure what can be regarded as appropriate attributions of responsibility.

However, although Watson argues that our human nature constrains the appropriateness clause and makes it unlikely for us to hold very young children responsible, it still leaves room for such situations to occur. It is not inconceivable to imagine cases in which someone may even demand goodwill from children. Hence, the problem of appropriateness is not entirely removed.

### 3.1.3 Michael McKenna's Moral Conversation

McKenna (2012) develops a distinctive understanding of moral responsibility, asserting that neither being nor holding responsibility is explanatorily more fundamental than the other. McKenna takes moral responsibility to be similar to a conversation, where reactive attitudes represent a stage in such an unfolding conversation. Just as a conversation includes an initiation, followed by a response, which in turn develops into further responses, an exchange concerning moral responsibility also begins with an initiation: an action by an agent that displays the quality of their will. The responses consist of others holding this agent responsible through reactive attitudes. Then, the one held responsible may respond by offering an explanation of her behaviour or acknowledging her wrongdoing, which may in turn lead to further responses. McKenna argues that to start such a moral conversation, every party in this conversation should possess the capacity for membership in the moral community. To be held responsible, one need not be a member of the community. Furthermore, understanding moral demands to continue the conversation does not require adopting the shared framework of values. As long as one possesses awareness and understanding of what one's actions reveal about intentions and the quality of will, one can be morally responsible. Facts about the agent also play an explanatory role in the considerations of others. According to McKenna, individuals can engage in moral communication only if they understand what they are communicating, with this knowledge arising from the control they exert over their actions. Both considerations of holding and being responsible will lead to a successful moral responsibility exchange. In this framework, reactive attitudes are appropriate insofar as they are meaningful within the context of the conversation.

However, although McKenna claims that neither being nor holding responsible is explanatorily more basic than the other, he appear to have a preference. For example, he draws an analogy and argues that the activity of the electron explains its features. As he clearly states, "it is the electron that is more fundamental or basic in the order of being as in comparison with the goings on in the cloud chamber" (McKenna, 2012: 40). Although he contends that explanatory priority does not imply that one is metaphorically more fundamental, he appears to place the electron as more basic than its activities. Similarly, although holding responsible may have explanatory priority over being responsible, this does not entail that moral practices are metaphysically more basic. However, he appears to prioritise being responsible, which turns him away from Strawson's account and draws him back to the traditional debate.

### 3.2 Summary: Why these Strawsonian Approaches Fail?

As mentioned above, the naturalist turn is central to Strawson's account, which leads to a unique theory that transcends the optimist/pessimist dichotomy. For Strawson, it is our moral practices that determine, or at least partly determine, what it means to be responsible. However, I demonstrated that the developments of Strawson's account fail to capture its uniqueness. Both Wallace and McKenna face the threat of stepping back into the traditional moral responsibility/determinism debate. Watson does endorse the naturalistic turn, but his account still leaves room for a response from the pessimist. I think that the root of their problems may lie in their search for the conditions of appropriateness itself. What they have in common is that they attempt to balance the appropriateness clause with the naturalistic turn, seeking to identify moral capacities while adhering to Strawson's account. However, as previously mentioned, identifying moral capacities means grounding the justification of Strawson's account in conceptual analysis, seeking to theoretically determine the capabilities that qualify individuals as appropriate moral agents, rather than justifying moral behaviour through the application of everyday practices. Therefore, it is always open for the pessimist to offer capacity-friendly counterexamples as challenges. For example, they may acknowledge that the capacity to understand moral demands can determine the appropriateness of a moral agent, but they might still challenge: Yes, but what is the relationship between this capacity and the truth of determinism? What if the agent is manipulated to have this capacity? What about a super AI that can understand moral norms? The underlying logic of those so-called Strawsonians is to view moral practices such as reactive attitudes—the facts of moral responsibility and moral capacity—as stable and unchanging entities, merely awaiting theoretical content to be filled in, much like engaging in an intellectual game. Being responsible is fixed to its place, and holding responsible is also fixed in its place. These so-called Strawsonians fill in their own understanding of the nature of these conceptions, searching for conditions for moral responsibility accordingly, and, based on their interpretation of these concepts, combine and arrange them to construct the relationship between holding responsible and being responsible. When they posit conditions for when it is appropriate to hold people responsible, there appears to be little distinction between these Strawsonian approaches and traditional compatibilist approaches.

## 4. Two Possible Methods to Save the Strawsonian Approach

There may be two ways to respond to the threat of the pessimist and the naturalistic turn. One way, which I find philosophically promising, is to view Strawson's account from a perspective of a language-game and argue that there can be no meaningful question regarding the justification of moral practices. Since it is meaningless to pose such questions, the threats that Strawson's account is confronted with dissolve. Another way is to acknowledge the historical and processual dimension of moral considerations and investigate how they evolved from a primitive stage to acquiring various contents over time. Since most Strawsonians seem to construct castles in the air, indifferent to the dynamic development of these concepts, their attempts to fill these concepts with content are based solely on their own understanding of Strawson's account and their imagination of the overall moral responsibility landscape. Returning to the historical context[3] and considering moral practices and concepts as progressive changes will help us understand at which stage holding responsible and being responsible acquired specific meanings. It will also illuminate how these concepts naturally became intertwined during the process of historical development. Subsequently, I will briefly outline these methods, then integrate these methods with Nietzsche's texts and argue that Nietzsche's texts can be interpreted in a Strawsonian way, with Nietzsche's genealogical insight further developing Strawson's approach.

### 4.1 Language-Game

Audun Bengtson (2020) offers a reading of Strawson's naturalistic turn from the aspect of language, which I consider philosophically promising. Bengtson highlights our primitive reactions

[3] While the historical context of moral responsibility provides some insight into its origins, what truly matters in understanding moral responsibility is the process through which it evolves within that context. Rather than focusing on specific historical events, we should direct our attention to how moral responsibility gradually develops through stages of human growth and social interaction, all of which unfold in a dynamic historical environment.

and "very general facts of nature", concepts that resonate with Strawson's naturalism. According to Bengtson, these primitive reactions serve as the foundation for language's development. In other words, it is these primitive reactions that influence the structure of our language-game, which concerns the grammatical rules governing the use of words. For instance, when teaching children the meaning of the word "pain", we often rely on their primitive reactions to pain as samples to explain its meaning. In this process, children will get to know how to use this word in their lives as well as the meaning it represents. These primitive reactions, coupled with the environment we inhabit, form the most general facts about human beings, thereby conditioning the grammar of the words and concepts we use and are associated with these facts. Therefore, our conceptions are neither discovered in the world through analysis nor rationally constructed or justified. Rather, these conceptions are shaped both by our nature as human beings and the world we inhabit. Over time, they gradually evolved from our primitive reactions to others' behaviours and developed through natural forms of communication. Two implications can be drawn from the perspective of a language-game. First, since human nature and the context of the world we inhabit shape and form our conceptions and language-games, we have no choice but to participate in these games and develop the capacity for language as human beings. Regardless of the cultural context we find ourselves in, our language acquisition is rooted in primitive reactions to life experiences. While variations in pronunciations and grammar naturally emerge over time, the application of specific concepts in daily life remains consistent. Whether referring to "pain" or "痛," both denote an unpleasant sensation or emotion caused by illness, injury, or harmful external stimuli. In this way, the use of "pain" and "痛" in different cultural contexts is also associated with negative emotions. Therefore, we are inescapably committed to our language-game. The process of language acquisition is neither reasonable nor unreasonable, neither justified nor unjustified because our primitive reactions stems from our nature. The second implication is that if the facts of nature change, so too will the grammar and its meaning. For instance, consider the fact that "we dial 999 for the police". If this fact changes, and we instead dial 110 for the police, then our use of the word "police" will change, indicating that we are playing a different language-game.

What relates language-game naturalism to Strawson is understanding reactive attitudes as the primitive reactions of which our moral responsibility language-game is an extension. In the same way that primitive reactions, such as pain, communicate information about the state of pain, our reactive attitudes similarly communicate resentment in a simple way, expressing our disapproval of others' behaviour. The moral responsibility language-game evolves from this primitive form of communication and determines the meaning of words regarding moral responsibility, such as desert, blame, excuses, justifications, and so on. In other words, reactive attitudes, which are reflections of our basic concern and demand for goodwill as humans, determine how words are used and, consequently, their meaning and the grammar of our moral responsibility language-game. Thus, moral judgements are made based on facts about our nature and the world we inhabit. If the world or the very general facts of our nature change, then our moral judgements conditioned by these facts will also change.

Recall that Strawson's naturalistic turn faces the problem of holding very young children responsible. To avoid this issue, Strawson could posit conditions under which it is appropriate to hold someone responsible, which would draw him back into the traditional order of explanation where facts about the agent being responsible determine the practices of holding responsible. It would then be open for the pessimist to present counter-examples, and Strawson would only be able to respond by positing other compatibilism-friendly conditions. This would cause Strawson to fall into the traditional moral responsibility/determinism debate, which he seeks to avoid. However, the language-game interpretation of Strawson offers a potential solution to this dilemma. Our concept of moral responsibility and how we use our moral words are contingent upon what we are like as human beings and the world we inhabit. Given our nature and position as human beings, it is highly unlikely that we would view very young children as appropriate targets for moral attributions. To make it the case that very young children are seen to be responsible, we would need to acquire a set of general facts quite different from the ones we currently have. But once we do so, the concept of moral responsibility and the way we use moral words will be quite different as well, due to the change of our situation and basic human nature. In this way, we would engage in a different language-game, in which "being responsible" could take on a different use and meaning. In other words, if we were to begin feeling resentment towards very young children, this would not automatically imply that they are responsible within our current context, because what it means to be responsible in this imagined context would not be the same in our current understanding of responsibility. They are not responsible because their being responsible is not what we mean by being responsible. Therefore, the language-game interpretation of Strawson's account can resolve the problem he is confronted with. By linking our moral responsibility language-game to our general facts of nature as humans, Strawson can avoid identifying compatibilism-friendly capacities for being responsible, because he can now demonstrate that the tension between moral responsibility and determinism cannot be meaningfully formulated. To be more specific, when we ask whether very young children should be held responsible, we are already engaging in a different language-game from the one that concerns moral responsibility as we typically understand it. Therefore, the question itself is not valid within the current context. The language-game we are playing is a matter of our own nature as human beings, which is irrelevant to a deterministic world.

### 4.2 Borderline-style Theory

The language-game reading is philosophically promising. However, at first glance, this reading does not seem to have a direct connection to Nietzsche's texts. Later, I will explore the potential deeper link between the language-game interpretation and Nietzsche's work. For now, interpreting the second essay of *On the Genealogy of Morals* as tracing the developmental process of the conception of moral responsibility and the moral agent provides resources for an alternative resolution of Strawson's account—namely, the borderline-style theory. At the same time, Strawson's insights also inspire a new way of reading Nietzsche, which may, in turn, illuminate contemporary debates on moral responsibility. Nonetheless, the common ground between the language-game and the borderline-style theory lies in their view of the development of moral responsibility as a process. Here, I will briefly outline Strawson's general idea; in the next chapter, I will apply Strawson's framework to the reading of Nietzsche's texts.

Strawson does provide resources for dealing with the naturalistic turn problem, which are often overlooked by most interpreters. He considers the example of parenting children and claims that we often give compromised and borderline-style answers to the question "whether very young children should be an appropriate target of moral blame?" According to Strawson, children are "creatures that are potentially and increasingly capable both of holding, and being objects of, the full range of human and moral attitudes, but are not yet truly capable of either" (Strawson, 2008: 32). As a result, the treatment of children must contain a kind of compromise, which is situated between the normal reactive attitudes and the objective attitude. But what about the threat posed by determinism? Strawson claims that determinism does not allow for borderline-cases, because it would "be grotesque to think of the development of the child as a progressive or patchy emergence from an area in which its behaviour is in this sense determined into an area in which it isn't" (Ibid). To be more specific, consider how one would interact with her child. If her child makes her room a big mess, she would feel angry, but not the resentment she would have towards an adult. She would express her upset as a form of communication to let the child know that this kind of behaviour is not allowed, hoping that her child can learn from her mistakes. When her child grows older, she would be more responsible for cleaning up her own room. At that time, if the child still makes the room a complete mess every day, her mom will start to hold her responsible for what she did. In this situation, if the truth of determinism were to have any bearing upon the progressive modification of attitudes towards the child, this example would not be meaningfully formulated. The truth of determinism, if relevant at all, should apply consistently regardless of age. It would be absurd to claim that a child is not responsible before her tenth birthday because her actions are determined, but suddenly becomes responsible after turning ten because her actions are no longer determined. Determinism, if true, does not fluctuate with age, nor does age itself alter whether one's actions are determined or not. According to Strawson, the conditions determining the appropriateness of holding someone responsible evolve gradually over time. He highlights the key reasons why moral agents evolve gradually from inappropriate subjects of reactive attitudes to eventually becoming appropriate targets of moral attributions. Thus, Strawson's compatibilism suggests a straightforward and plausible response to the blameworthiness of young children. The threat of determinism posed by the pessimist is irrelevant to moral responsibility.

By examining Strawson's perspective on parenting and the progression of moral agents from childhood to adulthood, the borderline-style theory demonstrates that Strawson's compatibilism is attuned to the nuanced and gradual emergence of moral agents. The essence of Strawson's account lies in its emphasis on the intricate and multifaceted nature of our actual attitudes and practices of responsibility. According to Strawson, our understanding of children's responsibility must be grounded in the specific details of our attitudes and practices. To assert that responsibility includes more than that is to over-intellectualise this problem. If we believe there is more to responsibility than our attitudes and practices, we will be likely to overlook much of the richness and complexity inherent in our actual attitudes and practices. Strawson's perspective demands that discussions on moral responsibility seriously consider our degree-sensitive attitudes and practices of moral responsibility. Acknowledging the phenomenon of moral responsibility requires a genuine appreciation of its intricacies in human life.

Despite my rhetoric that children are potential responsible agents, I do not assert that the distinction between how we hold children responsible and how we hold adults responsible is merely a matter of degree rather than kind. It remains unclear whether the justification for holding children responsible differs fundamentally from, or is akin to, that for adults. Some may argue from a consequentialist perspective that the reason for holding someone responsible is to foster the cultivation of the addressee's moral agency. In the same way as adults who can have their moral agency cultivated by being held responsible, children's moral agency can also be affected by taking a nurturing stance, which is the most conductive way of moral agency cultivation for children.

Others may object that holding someone responsible is not solely for moral agency cultivation since, in real-life situations, we cannot predict how the other person will respond to our moral attributions. In an ideal situation, the person who is held to moral expectation would engage in a process of moral reflection and cultivate their moral agency, but in real life, faced with moral norms, individuals may react positively or negatively, agreeing or disagreeing, rendering the cultivation of moral agency uncertain. One might argue that managing this uncertainty is, to some extent, an inherent aspect of enforcing a moral norm: one must discern the position and potential reaction of the norm violator. By observing a person's response to accountability, one gains insights into their evaluative perspective and attitude toward themselves. This knowledge is crucial for establishing a framework for future engagements and interactions that promote positive outcomes for both the individual and others. If your friend deceives you and fails to recognise their mistake when confronted, it seems unlikely that they will cultivate their moral agency to become an honest person. Instead, this disagreement may reveal that they are untrustworthy, leading you to regard them as a fair-weather friend. In light of this framework, children are not responsible for the same matters as adults. We are unlikely to view children as those who can have robust disagreement with us and bear the consequence of moral disagreement. Even if a child disagrees with studying diligently, her disagreement may be merely based on personal preference rather than a comprehensive consideration of the potential social consequences. As such, we do not respond to the child's disagreement by reorganising our future interactions with them, such as giving up on them, but rather continue to educate and cultivate their moral agency. The differing ways in which we treat children and adults thus mark a qualitative difference in the justification and approach to their responsibility. When we hold adults to moral responsibility, we are open to the possibility of significant disagreement and reorganise our modes of interaction; when we hold children responsible, it is primarily for the purpose of fostering their moral education. Although both forms of holding responsible are grounded in forward-looking practices, we cannot treat adults and children in the same way.

Regardless of whether the difference in how we hold children responsible and adults responsible is due to variations in their respective degrees of moral development or the nature of our interactions with them, it is evident that children are not appropriate targets for reactive attitudes in the same way that adults are. Strawson's borderline-style case insightfully captures this distinction and offers a plausible resolution to the challenges introduced by the naturalistic turn. For whatever reason, children, as potential moral agents, occupy an intermediate position between the objective attitude and reactive attitudes. Unlike our approach to adults, wherein we assess their behaviour against moral expectations and modify our interactions with them based on robust responses, we perceive children as merely prospective moral agents to be guided, rendering our attitude towards them to some degree objective. However, the guidance is not imposed in a cold, detached manner. Regardless of whether we respond to a child's behaviour with composure or emotion, for most of us, our underlying expectation is to help them develop a sound moral agency. As is shown in the birthday example, the determinist position fails to account for the developmental trajectory through which children cultivate moral agency, as the conditions determining the appropriateness of holding someone responsible evolve gradually over time, while determinism lacks the capacity to accommodate the compromises inherent in guiding their moral growth. Determinism does not fluctuate with age. Thus, we possess the resources to elucidate why children should not be regarded as appropriate targets of reactive attitudes in the same manner as adults, without relying on conditions external to human agency.

Moreover, although they seem to be far from similar, the language-game reading and the borderline-style theory also have a common point. Both of them avoid constructing a static conceptual framework that begins with a broad theoretical outline and is subsequently refined through the addition or adjustment of specific concepts. Instead, they prioritise the dynamic developmental process of moral practices and moral agency. For the language-game reading, the language-game of moral responsibility develops from our basic human nature and gradually acquires its own grammar, thereby endowing moral conceptions with meaning. Similarly, the borderline-style theory positions children within the spectrum between an entirely objective stance and an absolutely participant stance in the development of moral agency and states how our reactive attitudes fluctuate corresponding to the developmental stages of moral agency. Therefore, taking a genealogical approach to trace the historical evolution of moral conceptions and examine the role of our reactive attitudes proves valuable for understanding moral responsibility. In the next chapter, I will argue that Nietzsche's works, particularly the second essay of *On the Genealogy of Morals*, enhance and substantiate both perspectives. Interpreting Nietzsche's genealogy of the historical development of moral responsibility through a Strawsonian lens not only clarifies the intricate complexity of human moral attitudes and practices by providing textual evidence to support these two theories, but also sheds light on a new reading of Nietzsche in the context of moral responsibility. This approach integrates Nietzsche's poetic yet fragmented reflections into a coherent and systematic framework and reveal the underlying psychological structure of taking responsibility. Through this method, Nietzsche's seemingly ambiguous and fragmented perspectives on moral responsibility and free will are synthesised into a coherent theoretical account.

It is true that Nietzsche affirms certain forms of morality while criticising others, even though both are described using the same terminology. As I will argue, these concepts of moral responsibility do not emerge independently of one another; rather, they are internally related and historically continuous. The form of moral responsibility Nietzsche endorses is the one that expresses self-worth through pride and courage in taking responsibility for one's actions. Yet under the influence of Christianity, this same capacity for responsibility can become distorted into a form of self-doubt that undermines one's sense of existential worth, ultimately resulting in a devaluation of life itself. Therefore, Nietzsche is not rejecting moral responsibility as such, but a particular expression of it—a corrupted form rooted in ressentiment and self-contempt. Strawson, crucially, provides the conceptual resources to analyse this psychological structure, clarifying that the sovereign individual is not an ahistorical invention but a figure who emerges from within the very practices of the morality of custom. In other words, the task is not to resolve an apparent contradiction in Nietzsche's texts, but to clarify the internal relationships between the concepts he deploys. By unpacking the psychological structure underlying the lofty terms Nietzsche uses—such as "autonomy," "self-mastery," and "responsibility"—we can recover the deeper coherence of his moral psychology. Strawson's framework offers precisely the interpretive tool needed for this task. In this way, questions such as why Nietzsche criticises traditional morality while praising the individual who assumes responsibility, what kind of moral responsibility Nietzsche advocates, how it emerges from his critique of traditional moral responsibility, and how it connects to free will, are rendered more intelligible through this practical perspective.

# Chapter III

# Conscience, Bad Conscience and Moralised Guilt

## 1. Introduction

As argued above, Strawson reverses the metaphysical interpretation of responsibility by inverting the traditional order of interpretation and questioning the relationship between being and holding responsible. This move by Strawson complements Nietzsche's rejection of metaphysical accounts of moral responsibility. While Nietzsche critiques such accounts for their implausibility and their close relationship with Christian religion, Strawson begins from the ordinary practices and affects of holding responsible rather than from metaphysical preconditions. Conventionally, the metaphysical interpretation seeks to justify responsibility by appealing to preconditions such as control and free will, irrespective of the practice of holding responsible. By contrast, Strawson argues that the justification for responsibility lies within the practice of holding responsible, which is expressed in our reactive attitudes and the two types of "pleas". Building on this foundation, he offers his central argument to show the inescapability of moral responsibility. While his argument require further development to be fully persuasive, what I aim to retain from Strawson is his general methodological framework: specifically, his psychological analysis of responsibility, which defines its nature and conditions in terms of our "reactive attitudes".

Strawson's argument seeks to mediate between the compatibilist and incompatibilist positions in the debate over moral responsibility, thereby responding to the challenge determinism poses to our practices of praise and blame. Nietzsche, by contrast, is not principally concerned with whether we can still be morally responsible if determinism is true. Although Nietzsche's primary target differs, the methodological similarity allows us to bring them into dialogue. Through a

Strawsonian lens, it becomes clear that Nietzsche, like Strawson in "Freedom and Resentment", reverses the traditional metaphysical interpretation of responsibility in *GM* II. That is to say, rather than approaching responsibility through a conceptual or metaphysical analysis of its relation to determinism, Nietzsche, too, examines it through the lens of the practice of holding oneself and others responsible—a practice that is grounded in what he calls the "reactive affects" (*GM* II:11), which closely resemble what Strawson terms the "reactive attitudes". Nietzsche seeks to offer a practical account of why human beings are regarded as responsible by tracing the developmental emergence of conscience, bad conscience, and guilt—psychological capacities plausibly understood as lying at the heart of moral agency. Specially, he traces how guilt, as a certain form of reactive attitude to misdeeds, originates from the human historical interplay of punishment, socialisation and the development of a non-moral conscience.

In this chapter, I aim to investigate what Nietzsche says about the concept of moral responsibility by connecting his ideas with Strawson's and exploring, through a Strawsonian reversal, Nietzsche's own interpretation of moral responsibility. Nietzsche explicitly articulates his reversal in the fourth aphorism of *GM* II. Here, he emphasises punishment as a form of moral practice and criticises earlier "genealogists of morals" for failing to recognise the concept of *Schuld* (guilt) as a reactive affect that arises in response to others' moral practices. He further defines punishment—central to the emergence of morality—as requital or retributive, claiming that past genealogists overlooked the fact that this practice "evolved quite independently of any presupposition concerning freedom or non-freedom of the will" (*GM* II: 4). Nietzsche's understanding of retributive punishment thus amounts to a "reversal" of the metaphysical interpretation of responsibility because he analyses this conception without taking into consideration any metaphysical belief such as alternative possibilities or free will, but rather in terms of the practices of holding others to certain expectations and the corresponding feeling of guilt. In the second essay, Nietzsche situates the emergence of guilt within the historical nexus of punishment, socialisation, and the formation of conscience. As I will argue, moral responsibility is a concept that evolves through moral practices, where the transformation of guilt reveals how responsibility can first attain a healthy form and later become moralised—a process Nietzsche ultimately opposes. In what follows, I will apply this Strawsonian lens to Nietzsche's genealogical account in *GM* II, showing how it illuminates the psychological construction of responsibility and clarifies what Nietzsche might ultimately mean by "being responsible".

The discussion begins with a customary reading that regards conscience as the product of the "creditor-debtor" (*GM* II: 5) relationship. Conscience, as a "memory" (*GM* I: 3), or the awareness of one's obligations, develops as a result of punishment. Punishment fixes rules and norms into one and creates a memory within the individual. Thus, the emergence of conscience makes us cautious of social rules and helps us avoid punishment. The question thus arises: what circumstances led to our punishment? In other words, under what conditions did conscience originate? A seemingly plausible interpretation suggests that conscience emerged within the contractual relationship between creditor and debtor (*GM* II: 5). According to this reading, conscience originated in order to track obligations through the repayment of promises. However, I argue that this interpretation is flawed, as it assumes people can communicate and make promises to enter into contractual relationships. However, conscience is a prerequisite for making promises. This interpretation begs the question, as it presupposes the existence of conscience.

In the second section, I propose an alternative reading that regards conscience as the foundation for making promises. To do so, I trace the genealogy of conscience, outlining how firstly humans become "calculable" and conform to norms, then how they develop conscience to align with social expectations and build contractual relationships, and finally how they transform prescriptive rules (Frans de Waal, 1991) into bad conscience and moralised obligation. I argue that primitive conscience takes the form of "I will nots" to maintain social existence and makes promise-making possible. Then conscience takes the principle of equivalence into account and transforms the failure to meet social expectations into a debt. Debt is turned inward, giving rise to bad conscience that makes feelings of guilt possible. Guilt eventually becomes tied to normative judgements about self-worth, transforming into a sin of which one is consciously aware. In this way, there is no inconsistency in Nietzsche's writings, as conscience precedes contractual debts rather than being a consequence of them. Nietzsche's perspective on moral responsibility—specifically, his endorsement of a healthy form of responsibility and rejection of moralised guilt—is revealed through his account of the historical transformations of this concept within practices. The new understanding of moral responsibility lays the foundation for the subsequent discussions on free will.

## 2. The Contract Reading vs The Rules Reading

### 2.1 The Contract Reading

The contract reading is the view that conscience is developed through a contractual relationship. In this view, the debtor makes a promise to the creditor for some purpose. After making this promise, the creditor holds the debtor to the expectation of repayment. Since the debtor has not yet developed conscience and the memory of her obligation, she fails to repay and is punished by the creditor as an alternative way of compensation. The punishment is not regulated by the principle of equivalence and does not aim to rectify the debtor's behaviour but serves to satisfy the creditor's "instinct for cruelty" (*GM* II: 5). To avoid punishment in the future, the debtor develops conscience as a memory of obligations to remember debts and maintain the motivation to fulfil promises.

Nietzsche appears to endorse the contract view in his text. For example, he argues that "these contract relationships admittedly awakens various kinds of  suspicion and resistance toward the earlier humanity that created or permitted them… Precisely here there are promises made; precisely here it is a matter of making a memory for the one who promises". In this phrase, what Nietzsche argues is that in making promises and "to provide a guarantee for the seriousness and the sacredness of his promise", the creditor punishes the debtor by inflicting cruelty and pain "to impress repayment on his conscience as a duty, as an obligation (*GM* II: 5)". In this way, the debtor develops conscience to recall her obligations. However, to build a contractual relationship, the debtor must be able to make promises for repayment before punishment is imposed. This creates the problem of begging the question.

Specifically, the contract reading is incoherent because it regards the act of making promises as a precondition for entering into a contractual relationship and for forming a subsequent memory of the debt. However, the memory of the debt, along with a sense of obligation, seems to be necessary for the very act of making promises. First, it must be emphasised that making promises involves more than merely communicating promises. Leiter argues that, even for Nietzsche's sovereign individual, the ability to make and keep promises is the "only" great task and represents a modest achievement. Although he concedes that it is not nothing, he expresses the thought that "it seems prima facie hard to square with the overblown rhetoric of *GM* II 2" (Leiter, 2019: 150).

However, what truly matters is not simply making promises, but the ability to keep promises. As Regnister puts it, "to make a promise is to commit to doing something at some appointed future time even if doing so has by then become contrary to my 'private desires and advantages' (*D* 9)" (2017: 4). By making promises, the debtor not only expresses her intention to enter into a relationship but also binds herself to that promise. Both the debtor and creditor understand repayment as non-optional, meaning that making promises is not a simple task of communicating intentions, but involves "an intention to undertake an obligation." Consequently, making promises requires the concept of obligation, which is communicated by conscience.

What does it mean to say that making promises requires that the debtor already has a concept of obligation? It is noteworthy that this obligation does not equate to moral obligation, nor does it declare the debtor guilty if they break the promise. In this context, the term "obligation" pertains only to the state wherein the debtor binds herself and remains committed to fulfilling the promises. In other words, she "should" or "ought to" do so, regardless of any external circumstances or personal desires she is confronted with. The requirement to meet the creditor's expectations is mandatory. If the debtor fails to keep her word, the creditor has the authority to impose cruelty on her. This depiction clearly illustrates that the debtor is considerably less likely to make promises without first possessing a notion of obligation, as dictated by social norms.

Moreover, the debtor must also possess conscience in order to recognise her debt and obligations. Nietzsche defined conscience as "a real memory of the will" (ein eigentliches Gedächtnis des Willens), which is not merely having a desire to do something, but rather "a desire for the continuance of something desired once" (ein Fort-und fortwollen des ein Mal Gewollten) (*GM* II: 1, Kaufmann trans. / *eKGWB*). Therefore, conscience serves as a motivational structure that makes us aware of our obligations, our "extraordinary privilege of *responsibility*" (*GM* II: 2), and binds us to our promises. Without conscience, the debtor would forget not only her obligations but also the very concept of obligation itself, rendering her incapable of making promises.

The whole process is mapped out below:

(Conscience) → Making promises → Contractual relationship → Conscience

To summarise, the contract reading creates a contradiction between these two propositions:

A. A contractual relationship is formed by making promises, which requires the existence of conscience.
B. Conscience originates as the memory of obligations stemming from a contractual relationship.

To develop the memory of obligations as a sense of debt, the debtor must first be capable of establishing the contractual relationship. To establish the contractual relationship, the debtor must be able to make promises, which presupposes that she already possesses the obligation of repayment. Since, in the contract reading, the debtor lacks a concept of obligation prior to entering into the contractual relationship, it would be impossible for her to form such a contractual relationship. Nietzsche's account of the origin of conscience seems to encounter a contradiction.

I do not regard this contradiction as an insoluble problem. As mentioned above, what Nietzsche refers to as making promises and obligations are merely non-moral obligations, or more precisely, pre-moral obligations at this stage, which are the initial phrases of the moralised notions that follow. In this regard, Nietzsche would accept proposition A while rejecting proposition B, since making promises requires only a primitive form of conscience. In the next section, I will explore the developmental stages of these initial obligations and examine how they come to constitute voluntary debt.

### 2.2 The Möbius Reading

As previously mentioned, the contract reading regards promise-making as the precondition of a contractual relationship, wherein transgressions result in punishment and the memory of obligations. In this way, Nietzsche would view conscience as an outcome of contractual relationships, because the fulfilment of contractual obligations and social interactions can shape one's consciousness of certain moral and legal standards. However, a problem arises from the fact that entering into a contractual relationship requires that one already has a concept of obligation. Thus, I propose an alternative reading, suggesting that the relationship between conscience and promise-making resembles a Möbius strip, where they intertwine and mutually influence each other. Conscience plays a role in promises, and contributes to the contractual relationship, which in turn shapes a more advanced notion of conscience.

To explore this reading in more depth, it is necessary to determine whether conscience is indeed a prerequisite for making promises. As Nietzsche writes in the opening of the second essay, promising requires one to be "calculable, regular, necessary" and to "stand security for *his own future*" (*GM* II: 1). In other words, one who is able to make promises is "calculable" and "regular" in the sense that she is able to discriminate between obligatory and non-obligatory behaviours, so as to conform to social norms rather than act solely on sudden and strongest desires. Nietzsche traces the history of conscience to explain how humans become "regular" in this way. For him, conscience is a "*memory of the will*" (Ibid.) that helps humans "'remember' social rules or practices, to be bound by them" (Richardson, 2004: 89).

Some may cast doubt on the possible existence of a comprehensive concept of obligation prior to the establishment of a contractual relationship, since it seems implausible for conscience to emerge without the experience of entering into binding agreements with others and suffering the consequences of failing to honour them. However, this reading regards conscience as unchangeable, which I thinks is not promising. According to Nietzsche, the conscience of the sovereign individual exists in its highest form and "has a long history and variety of forms behind it" (*GM* II: 3). This suggests that conscience is not a static concept but is constantly evolving, progressing from a basic memory of rules to its highest form. In other words, conscience is a concept that can be distinguished by degrees. For the act of making promises, the conscience does not need to be a complex moralised concept but simply consists of a memory of certain social rules that humans intuitively recognise as binding for social cooperation. These are basic guidelines that members of society feel compelled to follow, such as "I will not kill", "I will not rob", "I will not cheat" and "I shall do this". As Nietzsche puts it, to cultivate the ability to make promises is itself "the long story of how *responsibility* originated", which presupposes conscience "as a preparatory task" (*GM* II: 2). The conscience required for entering into contractual relationships, therefore, is merely an early stage in the development of responsibility, a stage that can be traced back to our evolutionary past. It is thus reasonable to suppose that there exists a primitive form of conscience, one that acts as the precondition for making promises.

Nietzsche's view on conscience echoes the perspectives of many contemporary ethologists and anthropologists. For instance, Christopher Boehm (2010) traces the evolutionary origins of human morality back to the earliest human societies and notes that despite the inherent selfishness, aggression, and nepotism observed in humans' primate counterparts, their social behaviours were governed by implicit rules and norms of ancestral human communities for the greater benefit of the group as a whole. As he writes, "No matter what size the society, people everywhere seem to realise that by reinforcing and amplifying individuals' tendencies to extrafamilial generosity, they can improve the overall efficiency of cooperation from which everyone profits. At the same time, they understand that failure to reciprocate can cause conflicts that seriously disturb group social or economic life" (Boehm, 2010: 51). These constraints were not yet moralised concepts and served merely to avoid trouble, rather than incite a sense of guilt. However, it is important to note that primates may also display positive social behaviours such as play, care for the elderly, and cooperation, which contribute to the social cohesion and well-being of the group. Frans de Waal writes that "In human behaviour, there exists a tight relation between empathy and sympathy, and their expression in psychological altruism. It is reasonable to assume that the altruistic and caring responses of other animals, especially mammals, rest on similar mechanisms (de Waal, 2006: 28). According to de Waal, humans, along with other species such as chimpanzees, possess empathy and sympathy toward others. These emotions "emerge at an early age in virtually every member of our species" (Ibid.), therefore forming the basis of conscience and prevent us from engaging in certain actions. Whether the evolutionary processes began with collaborative action for survival or the expression of empathy and sympathy, humans and animals like monkeys and chimpanzees gradually came to conform to specific social rules beneficial to their aims. What makes humans "calculable" are precisely these kinds of rules that generate a sense of obligation, as Nietzsche claims.

How did this primitive conscience originate? How did humans come to act on the basis of primitive conscience? Many believe that conscience generates obligations because it is a learned behavioural pattern actively enforced through reward and punishment. For example, supporters of the contract reading believe that the contractual relationship entails the principle of punishing those who have transgressed the rules they are supposed to obey and rewarding those who meet expectations. The pain of punishment would make people in the contractual relationship fearful, encouraging them to develop a consciousness of social norms in order to avoid punishment. Nietzsche acknowledges that the contractual relationship and the principle of reward and punishment it entails indeed contributed to the development of conscience, as "pain is the most powerful aid" (*GM* II: 3) in creating memories that fostered a tendency to follow rules. Living in hierarchical relationships, humans were subjected to the authority of those in power. They developed the memory of following rules due to fear of being expelled from society and fear of those in power. However, memories developed in this way can only create a mechanism for following rules, rather than a genuine recognition of the obligation to follow rules for their own sake. In *Daybreak*, Nietzsche argues that conformity to customs requires obeying the rules "not on account of the useful consequences it may have for the individual, but so that the hegemony of custom, tradition, shall be made evident in despite of the private desires and advantages of the individual" (*D* 9). Conformity to rules should not be, as de Waal says, motivated by the fear of punishment to minimise negative consequences, but should come from within as obligatory and unconditional. As a result, the contract reading has no way to account for Nietzsche's divergent views presented in his text. The vital importance of conforming to rules must lie elsewhere.

Nietzsche inadvertently reveals the answer when he regards these "I will nots" as "primitive demands of social existence" (*GM* II: 3). What constitutes society as a co-operative community is that humans living in it conform to social rules. In other words, rules must be followed in order to live as a social member among others. Obeying rules is thus obligatory because the co-operative society will not exist without doing so. As a notion that evolves with the concept of conscience, which I will investigate in detail in the following sections, punishment at this stage is merely an emotional release that deepens the memory of the need to follow rules due to the fear and excitement of the emotional outburst. Through this emotional outburst, people gradually realised that violating rules was detrimental to social stability and that following rules should be obligatory and unconditional, unrelated to the concepts of promises, contracts, or repayment linked to personal gains or losses. Even though noncompliance with social norms does not necessarily result in punishment, these norms must still be followed because adherence to them is an obligatory and non-negotiable precondition for the constitution of society. Therefore, obligation originated not only from the social fact of being obligated by those in power or authority to follow certain expectations and rules, but also encompassed a distinct mental state that involves recognising the course of action one "ought" to undertake, irrespective of personal preferences. These "I will nots" are simple social facts[4] that should be obeyed to live with others.

One point needs clarification. Although Nietzsche thinks that conscience originates as a memory of obligation to remind us of conformity to rules, it is not unconditional in the sense of Kant's categorical imperative. A key difference is that Kant's imperative is strictly moral, whereas

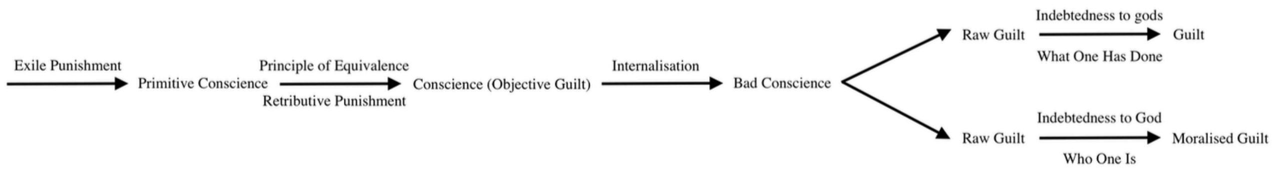
[4] By "simple," I do not mean that these rules are easy to implement or inherently simple in content; rather, I mean that these rules are not the result of careful design but exist as necessary conditions for maintaining society.

conscience, at this stage, does not yet involve the moral conceptions of obligation. Moreover, according to Kant, the categorical imperative is unconditional because it derives its authority from the rational impulse to obey the dictates of *Reason* itself. The categorical imperative commands absolutely without reference to any purposes and is therefore not subject to any variations or changes. As I argued in the previous chapter, Nietzsche denies the existence of a metaphysical self detached from affects and regards the pure moral motive as mere fiction. Yet, in *Daybreak*, he also claims that acting out of obligation should be regardless of private desires and advantages of the individual, which seems to bear some similarity to the Kantian notion. However, to reconcile the dilemma concerning "unconditional", Nietzsche offers a naturalistic explanation of the categorical force of social rules. For Nietzsche, the notion of obligation is neither dependent on nor totally independent from one's desires. Since the need for conformity "is now innate in the average man, as a kind of *formal conscience* that commands" (*BGE* 199), acting out of conscience is unconditional, regardless of desires and other factors other than conscience itself. Yet, this does not imply that conscience can show its categorical power only in the absence of one's affects and desires. It simply means that an "ought" does not depend on desires in order to form its existence. What it requires is "the backing of the rule" (Foot, 1972: 309), which does not necessitate a pure motive from which obligation develops. In fact, adherence to rules does not need to be pure. It is conceivable that fear and other emotions may enhance adherence to rules, norms and many "oughts", but one will not comply with them merely because of the possible outcomes they may bring. To obey unconditionally does not require adherence to reason alone, as in Kant's categorical imperative, devoid of any external influence. Instead, it simply requires regarding that one's conformity as it were, by itself "ought" to be done because it is a constitutive part of being a social member and maintaining society. In this way, emotions such as fear or concerns about the outcomes may influence rule-following, but conscience still gains deliberative priority in the complex real world, rather than in a metaphysical world.

So far, the relationship between conscience and promise-making has become clearer. The relationship seems like a Möbius strip, as they are intertwined and difficult to explain straightforwardly regarding their temporal sequence. Conscience, as a dynamic concept with various degrees, derives its obligatory power from the basic requirements of maintaining society. At this primitive stage in a long history, everyone was expected to comply with norms, and those who violated them would be punished. Punishment was merely an emotional release that deepened people's memories and helped to make them perceive following norms with a sense of obligation, enabling them to make promises and enter into contractual relationships. In a contractual relationship, punishment was no longer used as an emotional release but as a warning that violating norms is punishable. With the concept of "desert", conscience was continually strengthened in contractual relationships by the principles of reward and punishment it encompasses. Through numerous experiences of pain and cruelty, conscience ultimately reached its highest form as the ripe fruit of the morality of custom, enabling one to "stand security for oneself and to do so with pride" (*GM* II: 3). Therefore, Nietzsche can both view conscience as arising from pain and recognise its obligatory power of conscience as detached from affects and desires.

To summarise, Nietzsche suggests that although promise-making requires that one already has a sense of obligation, it could be a basic form of consciousness of prescriptive rules that originates at the beginning of a long human history. Conformity to "oughts" and "I will nots" is not merely about obeying rules and regarding following those in power as "unconditional" and, rather, about considering conscience itself as "unconditional". Humans became "calculable, regular, necessary" by developing this primitive conscience and subsequently made promises to enter into contractual relationships. From this perspective, the problem of the contract reading can be solved, as primitive conscience can be seen as an awareness of social norms and a sense of obligation, while promise-making is a behavioural pattern that further develops from this awareness and contributes to the highest form of conscience. The result of this development is the beginning of bad conscience, which I will discuss below.

## 3. The Development Process




### 3.1 Conscience, Punishment and Indebtedness

As previously discussed, Nietzsche views conscience as a memory warding against "forgetfulness" (*GM* II: 1) that originated for "the advantages of society" (*GM* II: 3). Primitive conscience served merely as a basic element of society, fostering an awareness of social rules and reinforcing adherence to norms, thereby producing "calculable", "regular", and "necessary" behaviours. However, Nietzsche observes that conscience in its primitive stage was just a basic element of society and, while not yet synonymous with moral agency, played a crucial role in its emergence. The function of primitive conscience was to sustain attentiveness to various "I will nots" arising from an intrinsic need to maintain society and be a part of it. As I will argue later in this chapter, Nietzsche suggests in *GM* that a norm attains moral status only when an individual is held to an expectation to follow it and experiences guilt upon its violation. At the primitive stage, transgression of norms led solely to severe punishment, often through violence and hostility, rather than evoking guilt in the transgressors or moral condemnation from others. Therefore, primitive conscience is neither sufficient for morality nor is it characterised by moral elements.

While initially functioning as a mere repository of social norms, primitive conscience gradually evolved to incorporate elements of indebtedness, shaping the practice of punishment in accordance with norms such as justice and fairness and thereby transforming into a form of conscience that began to include moral factors. Once the practice of punishment began to encompass these standards consented to by the whole community, it developed according to the principle of equivalence, which asserts that "every injury has its *equivalent* in something and can really be paid off" (*GM* II: 4). For Nietzsche, the idea that the creditor could punish the debtor as compensation for unpaid debts represents the "oldest and most naïve moral canon of *justice*" (*GM* II: 8), which transformed primitive conscience into "bad conscience" and later gave rise to a sense of guilt. Nietzsche does not provide an clear argument for this transformation; however, examining the impact of indebtedness on the practice of punishment is necessary to understand how primitive conscience developed into bad conscience.

In his second essay, Nietzsche distinguishes between two types of creditor-debtor relationships: the communal relationship between individuals and society, and the contractual relationship between individuals. These relationships correspond to two forms of punishment. In *GM* II: 9–10, he focuses on exile as a form of punishment at the social level, while in *GM* II: 5, he examines retributive punishment between individuals. Both types of punishment will be considered in turn.

### 3.1.1 Exile Punishment

Exile punishment pertained to the early stage of conscience, where the primary concern was to maintain social existence through the relationship between individuals and society. According to Nietzsche, in this relationship, the community stood towards individuals like a creditor to its debtor. An individual adhered to social norms and rules in exchange for protection and enjoyed "the advantages of a communality" (*GM* II: 9). The community did not offer advantages based on promises and contracts because obeying social norms was a basic social fact as an obligatory rule that did not require individual consent. Given the absence of communal standards to address transgressors or account for unpaid debts, punishment at this stage was the most severe. As Nietzsche puts it, "At this level of civilisation 'punishment' is simply the copy, the *mimus* of normal behaviour towards the hated, disarmed, defeated enemy, who has forfeited not only every right and protection, but also every mercy" (Ibid.).

The purpose of primitive punishment was to disgrace the transgressor like a defeated enemy in order to achieve various objectives. It enabled the entire society, particularly those who had been wronged, to express their anger and hostility towards the transgressors. It transformed transgressors into warnings and reminded members of the community of the consequences of breaking rules. However, the most important function was to banish the transgressor and cast them back "into the savage and outlaw state" (Ibid.) from which they had previously been protected. In other words, "the malefactor" (*GM* II: 10) violated the proto-contract with society and was thus expelled from the community. What happened subsequent to the act of exiling was nothing more than the severe and cruel venting of violence, because the malefactor's behaviours were "dangerous and destructive to the whole", which angered members of the community. Exile punishment, therefore, serves as a warning to deter potential violators, rather than seeking to compensate for erroneous behaviours. Punishment was conducted without standards, such as justice, fairness and desert. To avoid enduring exile punishment and being cast out of the community, individuals developed primitive conscience to remind them of their obligations for social maintenance.

$\xrightarrow{\text{Exile Punishment}}$ Primitive Conscience

### 3.1.2 Retributive Punishment

Retributive punishment pertains to relationships between individuals, rather than between a community and an individual. The shift from exile to retributive punishment occurred when society was stable and safe enough for individuals to engage in promises and contracts with one another. According to Nietzsche, the community grew in power and ceased "to take the individual's transgressions so seriously" (*GM* II: 10). Specifically, punishment was implemented following the principle of equivalence, and transgressors were no longer banished from the community; instead, they were given the opportunity to make amends and repair their relationship with those to whom they had made promises, and remain members of society.

As previously mentioned, primitive punishment was initially a random outburst of anger that was subject to no standards. When someone violated social rules, society, as a creditor, "could inflict every kind of indignity and torture upon the body of the debtor". For example, the creditor could cut parts of the body from the debtor "as much as seemed commensurate with the size of the debt" to satisfy the "enjoyment of violation" (*GM* II: 5). Nevertheless, when people reached a point where they could form contractual relationships within a relatively stable society, the intense anger directed at transgressors was "held in check and modified by the idea that every injury has its *equivalent* and can actually be paid back" (*GM* II: 4). In particular, humans were accustomed to "setting prices, determining values, contriving equivalences, exchanging" (*GM* II: 8), believing that debts can be measured and that creditors can request reimbursement from debtors equal to their debts. Consequently, community members sought redress for their losses as the primary concern, rather than a desire for cruelty, regardless of how transgressors were treated. As society grew stronger, punishment appeared more moderate and humane.

Furthermore, Nietzsche argues that the community began to assess crimes not solely on the basis of injuries they caused, but also considered whether they are "intentional," "negligent," "accidental," "accountable" (*GM* II: 4). In primitive stages, criminals were punished irrespective of whether they intended to break the rules or were reasonable enough to control their actions. At this stage, punishment served as a means of venting anger against the criminal's behaviour and intentions. The shift to isolating the criminal from their actions signifies a developmental transition, where punishment had developed beyond a mere venting of emotions to incorporate adherence to normative standards, such as fairness and justice, which were accepted by the community as whole. A person would no longer be expelled from the community simply because of their actions; instead, they would be judged based on their intentions and specific circumstances, allowing them the opportunity to make reparations, restore their relationships with others, and remain a part of society.

$$\xrightarrow{\text{Exile Punishment}} \text{Primitive Conscience} \xrightarrow[\text{Retributive Punishment}]{\text{Principle of Equivalence}} \text{Conscience (Objective Guilt)}$$

### 3.1.3 Reactive Attitude

Unlike Strawson, Nietzsche does not explicitly argue that free will and moral responsibility lie within the web of human relationships. While Strawson contends that the web of human relationships plays a crucial role in holding individuals responsible, Nietzsche does not provide a systematic argument for the connection between reactive attitudes and holding others responsible. However, he acknowledges that feeling guilty is part of the process of moralisation. Nietzsche argues that "punishment is supposed to possess the value of awakening the *feeling of guilt* in the guilty person" (*GM* II: 14). A rule qualifies as moral only if an individual is held to the expectation of obeying it, with the precondition that they would feel guilty for their transgression of it. From this perspective, Nietzsche appears to explore the role of our reactive emotions[5] in constructing morality when rules are violated in society.

The practice of exile punishment involves directing reactive attitudes, such as anger and resentment, toward criminals who have caused injuries, as punishment for their transgressions. In primitive society, humans were expected to conform to social rules for the maintenance of society and were regulated by those in authority. Those who violated these rules received punishment as a means of venting anger and resentment, as they had broken the expectations imposed upon them. However, when punishment incorporated the principle of equivalence and also took one's intention into account, the practice of holding others responsible and adopting reactive attitudes was no longer driven solely by emotions or dispositions to do so, but by a sense of appropriateness to see whether they deserved punishment and how they could repay their debts. In other words, reactive attitudes became measured according to norms that are agreed upon by society as a whole, such as fairness, justice, and desert. Mere conformity to rules no longer constituted the core of holding individuals responsible. As Clark states it, "With the belief that the infliction of pain and suffering is justified, that the offender owes or deserves it, it seems that we have a primitive case of holding

[5] Nietzsche appears to imply a broader range of reactive emotions than Strawson because, as demonstrated here, he considers emotions like raw guilt—which is not yet a moral notion—to be part of what gives rise to moral responsibility. In contrast, Strawson limits reactive attitudes to moral practices, such as gratitude and resentment.

responsible or accountable and of blaming behaviour, and one that involves no incompatibilist assumptions" (2015b: 95).

While it is clear how individuals are held responsible for their behaviour, it remains ambiguous how they can hold *themselves* responsible. The notion of guilt at this stage is purely objective, indicating a material sense of debt rather than a notion intrinsic to one's self-worth. As Greenspan (1992) explains, objective guilt pertains to external factors such as laws, rules, and social norms that define right and wrong. It refers to situations in which an individual has committed a wrong or violated a normative standard. In this sense, one who violates social rules would be considered guilty in accordance with social norms, both by herself and others, but she may not feel the subjective feeling of guilt or the need to apologise for her wrongdoings. Nietzsche distinguishes between subjective and objective guilt when comparing the punishment of malefactors to the punishment of children. He argues that throughout most of human history, "punishment was *not* imposed *because* one held the wrong-doer responsible for his deed, thus *not* on the presupposition that only the guilty one should be punished: rather, as parents still punish their children, from anger at some harm or injury, vented on the one who caused it" (*GM* II: 4). Children are not morally developed to feel guilty for violating rules, but they may be blamed by their parents for doing so.

Despite the feeling of remorse and the willingness to apologise, the legal system, due to its nature of procedural punishment for crimes without involving any intrinsic emotional factors, fails to shame offenders through its judicial procedures and imprisonment. It merely makes criminals aware that powerful people are permitted to do things that they are not allowed to do. Offenders understand that something has gone wrong and that their actions fall short of their expectations, but they do not recognise their own wrongdoings. As a result, while punishment may tame human behaviour, it does little to promote guilt, remorse, and self-accountability. Punishment and its resulting objective guilt fulfil the needs of the community but fail to demand accountability from offenders themselves.

Subject guilt arises only when individuals internalise objective guilt, recognising not only their indebtedness but also feeling that they deserve punishment for violating rules. As Clark observes, "It is only when one internalises (i.e., adopts against oneself) the hostile attitude of one who thinks you owe him something, and, more specifically, that you deserve to suffer for what you owe him, that it starts to be recognisable as guilt" (2015a: 70). In the following section, I will expound on the development of the bad conscience and how it gave rise to the shift from objective to subjective guilt.

**3.2 Internalisation and Holding Oneself to Expectations**

Although punishment can cultivate habits of conformity to norms, which Nietzsche calls the "instinct of obedience" (*BGE* 199), he maintains that morality is not solely about compliance with rules. Individuals may recognise that they have violated social expectations, done wrong, and deserve punishment—yet, they are unable to feel guilty for their behaviour. Therefore, Nietzsche believes that morality requires not only adherence to norms but also the ability to experience guilt when one breaks them. Only then can they regard adherence to norms as binding, connect it to their self-worth as a good person, and hold themselves accountable when violating social norms without the requirement of others' blame and punishment as motivation.

Consistent with my interpretation that guilt has a long history and developed from primitive conscience, it is plausible to view the ability to feel guilt as not innate but rather as the result of the emergence of bad conscience. To feel guilt, individuals must experience internal pain and remorse, or they must undergo the "pang of conscience" (*GM* II: 14). This feeling is then interpreted as a moral failure compared to the standard one accepts for a morally good person. In this section, I will focus on how the internal pain, "the actual *instrumentum* of that psychical reaction called bad conscience" (Ibid.), is generated.

### 3.2.1 Internalisation

According to Nietzsche, bad conscience is not the same as primitive conscience. Rather, it is a pang of internalised aggressions that engender guilt feelings. According to Nietzsche, this bad conscience arises from a process called "internalisation":

> All instincts that do not discharge themselves outwardly tum inward—this is what I call the internalisation of man…Those fearful bulwarks with which the political organisation protected itself against the old instincts of freedom—punishments belong among these bulwarks—brought about that all those instincts of wild, free, prowling man turned backward against man himself. Hostility, cruelty, joy in persecuting, in attacking, in change, in destruction—all this turned against the possessors of such instincts: that is the origin of the "bad conscience." (*GM* II: 16)

During the early phase, guilt triggered by the internalisation of aggressions was described as being in a "raw state" (*GM* III: 20). Initially, while attacks on one's own instincts caused inner pain, they did not necessarily indicate that individuals were aware of their wrongdoing or consider their behaviour as failing to meet moral expectations.

Why do humans turned their aggressions inward, especially given that they used to direct their hostility and anger towards others as punishment for violating rules? According to Nietzsche, the peaceful environment of modern society rendered all human instincts "disvalued" and "suspended" (*GM* II: 16). To sustain society, humans were compelled to suppress and reframe their tendencies for cruelty and were reduced to "thinking, inferring, reckoning, co-ordinating cause and effect" (Ibid.), rather than relying on unconscious drives. As Nietzsche posits in the next aphorism, internalisation of aggression did not stem from the typical process of socialisation but arose from "acts of violence" (*GM* II: 17). He explains that the origin of such violence occurs when "a conqueror and master race which, organised for war and with the ability to organise, unhesitatingly lays its terrible claws upon a populace perhaps tremendously superior in numbers but still formless and nomad" (Ibid.). The conquerors introduced strict laws for social command and for domesticating the conquered, resulting in an environment where people could no longer freely express their instincts for cruelty and freedom. As a result, they had to suppress these impulses and turn them inward to torture themselves, which gave birth to the bad conscience. As Nietzsche writes, "This *instinct for freedom* forcibly made latent—we have seen it already—this instinct for freedom pushed back and repressed, incarcerated within and finally able to discharge and vent itself only on itself: that, and that alone, is what the *bad conscience* is in its beginnings". (*GM* II: 17)

It is worth noting that bad conscience and resultant raw guilt are universal in humans, as it is rare to conceive of an individual who does not feel regret or remorse for wrongdoing. Raw guilt also varies in degrees, as some individuals are more sensitive to their actions and more prone to punishing themselves when they think they deserve it. For example, victims of pick-up artist (PUA) relationships are more likely to hold themselves responsible for their unfortunate situation, while unfaithful partners may not regret their infidelity until they face serious consequences such as divorce. As such, bad conscience should not be considered an exclusively Christian phenomenon, nor should it be seen an inexpiable state as Christianity believes. Rather, it is an underlying mechanism that temporarily induces feelings of raw guilt.

### 3.2.2 Holding Oneself to Expectations

The concept of raw guilt is central to Nietzsche's theory of moral development. As raw guilt is the result of aggression directed towards oneself, it can be viewed as a form of self-punishment. When an individual suffered harm, punishment served as a reactive attitude towards the injury she suffered and toke the malefactor as the object of venting anger to avoid further harm. It could also alleviate negative affects suffered from being injured by expressing anger outward. However, under conditions of oppression, the capacity to release negative affects outward were constrained. Individuals might still release their natural instincts, but only by redirecting them towards themselves, making themselves the targets of their own anger. This, in turn, caused further harm and resulted in the desire to inflict suffering in response to the injuries sustained. Consequently, this process gave rise to an inner conflict whereby "an animal soul turned against itself, taking sides against itself" (*GM* II: 16), contributing to "the result of a forcible sundering from his animal past" (Ibid.).

The separation from the animalistic past has led to the fact that humans no longer directed their hostilities towards external enemies and resistances, but rather towards their internal enemies, which were their own natural instincts, "their regulating, unconscious and infallible drives" (Ibid.). According to Nietzsche, the "war against the old instincts" deepened the soul. In the past, when individuals were violated, their first reactive attitudes towards their injuries were to seek revenge and externalise their hostilities as compensation for the harm they had suffered. However, lacking the outlet to express their anger externally, humans found themselves unable to make the world yield to their will and were unwilling to accept this impotence. The only way for them to restore power was to revalue their values by turning their anger inwards, which transformed their desire for revenge itself into the object of their anger. In other words, individuals could only gain power over the internal world of their psychological structure by requiring that they denied their desire for revenge and developed a second-order desire to prevent their first-order desire from manifesting. In this way, our "entire inner world, originally as thin as if it were stretched between two membranes, expanded and extended itself, acquired depth, breadth, and height" (Ibid). Therefore, humans were no longer mere vessels of their desires and natural instincts but possessed a complex and extensive inner world.

Furthermore, the development of second-order desire from bad conscience allowed individuals to reflect on and hold themselves accountable to expectations, marking a turning point in the process of cultivating moral agency. Recall that when aggressions were not internalised, one might still possess a primitive conscience to follow rules, but their compliance was driven solely by habit out of their remembrance of social customs. One remembered the rules and was aware of others' expectations, yet one was unable to hold herself accountable to one's own expectations, and as a result, lived under the scrutiny of others as an object of their expectations. Through the internalisation of aggression, one became separated from animal instincts and regarded raw guilt, the pain within, as a punishment for these instincts. In other words, individuals punished themselves due to their desire for power, which was pursued through releasing anger and seeking revenge, but they recognised the impossibility of mastering the circumstances of the external world. To gain power, they imposed a certain form on their own motivational psychology by holding themselves accountable for not surrendering to their natural instincts. They punished themselves solely because they had violated the expectations they set for themselves. Only in this way could they experience guilt and negative emotions towards themselves, rather than mere embarrassment about violating external rules. These negative affects continued to fuel their desire for revenge, perpetuating the cycle of self-inflicted pain, thus establishing an enduring state for all humans.

The internalisation of aggressive instincts serves as the first step towards the formation of moral agents. An individual who has undergone this process must be capable of holding themselves accountable to their own expectations and experiencing feelings of remorse when they fall short of these expectations. However, they do not perceive this feeling as a failure to comply with expectations, unlike a morally good person would. In other words, they do not judge their anger, their first-order desire for hostility, as bad or evil. They may accept judgements concerning their guilt from both themselves and others, but they cannot identify with the judgement that they are guilty. Another aspect of being a moral agent, who can feel guilty as "the feeling of personal obligation", entails assigning value judgements to compliance with expectations.

$$\text{Conscience (Objective Guilt)} \xrightarrow{\text{Internalisation}} \text{Bad Conscience} \longrightarrow \text{Raw Guilt}$$

**3.3 Guilt with Value Judgement**

I have examined the concept of bad conscience as a mechanism for generating guilt in its raw form, which is not yet the feeling of guilt. As Janaway suggests, "there are plenty of imaginable ways in which such a mechanism could work without the suffering's being the specific one of feeling *guilty"* (Janaway, 2007: 133). For example, individuals might experience fear, shame, jealousy, or anger as a consequence of self-persecution within a socialised community. It cannot be definitively established that they "must feel guilt, have guilt consciousness, or indeed that they must suffer the pangs of bad conscience" even if they "must gratify themselves through self-inflicted suffering" (Ibid.). Clark similarly argues that guilt differs from raw guilt because guilt involves not only "the relation of the individual to someone she has injured or failed in some way", which is inherent in raw guilt, but also "the self's relation to the standards she accepts for a good person" (Clark 2001: 58). Clark contends that guilt encompasses additional elements based on raw guilt. Specifically, guilt entails a judgement about oneself and requires the acceptance of personal obligations as a morally good person, thereby linking judgement to one's self-worth. In other words, an obligation qualifies as moral only if violating it leads to a feeling of guilt, and this obligation becomes attached to guilt because it becomes personal and engages one's self-worth. As Janaway concludes, "What differentiates the feeling of guilt from other kinds of psychological pain… must be the way the subject represents herself: she must at least take herself to have done harm, to have transgressed, usually against some other agent, in such a way as to violate an obligation she accepts herself to be under (Janaway, 2007: 136). To understand how raw guilt was transformed into a subjective feeling one identifies oneself with requires an examination of how value judgements became attached to internal emotions and how one's failure to fulfil expectations was interpreted as "the feeling of personal obligation" (*GM* II: 8).

The origin of guilt presents a problem because Nietzsche does not provide a clear explanation of how and when raw guilt becomes linked to value judgements that one accepts in order to be a morally upright individual. As Janaway observes, it is something "Nietzsche does not explicitly provide for in his analysis—but which must be there nevertheless for guilt to occur—is the conception of oneself as a transgressor in one's own eyes" (Janaway, 2007: 136). While Nietzsche does outline how conscience develops from a memory of rules into bad conscience, which then generated raw guilt, this guilt remained in an intermediate state—neither fully objective nor subjective. The individual recognised their indebtedness to their own expectations but did not yet connect this recognition to their sense of self-worth. The key question, then, is how the material sense of indebtedness to one's expectations, as Nietzsche refers to (*GM* II: 4), transformed into subjective guilt?

Some may refer to this process as what Nietzsche calls "moralisation" (*GM* II: 21). It is worth noting that, the process by which obligations became moral is not equivalent to the process of moralisation. In the next section, I will argue that "moralised" guilt, which is indebtedness to the Christian God, is a manifestation of extreme and perverse guilt. Moreover, moralised guilt appears to presuppose the existence of raw guilt. Nietzsche himself clearly points out that "the involvement of the *bad* conscience with the concept of god" is subsequent to "the connection of the concepts 'guilt' and 'duty' with religious presuppositions" (Ibid.), where guilt was already in place. As Nietzsche notes, "the major moral concept *Schuld* [guilt] has its origin in the very material concept *Schulden* [debts]?" (*GM* II: 4), but moralised guilt was formed by "pushing back" debts (raw guilt) and the subsequent guilt "into the conscience" (*GM* II: 21) and developed from the awareness of one's failure to meet moral standards for interpreting one's sin as bad or evil. Here, I will trace our contractual relationship with ancestors, where Nietzsche claims that the "guilty feeling of indebtedness" (*Das Schuldgefühl*) already existed within the context of ancestor worship.

### 3.3.1 How Guilt becomes Subjective?

Although some argue that the consciousness of guilt arises from internalised natural instincts, it is important to note that the internalisation of aggressive drives is only one crucial component and is not identical to guilt feelings. The question then arises: how does objective guilt become attached to personal elements, giving rise to a sense of guilt? Earlier in the second essay Nietzsche offers an explanation for the origin of the feeling of guilt, which is sometimes overlooked. He states that "the feeling of guilt, of personal obligation, had its origin, as we saw, in the oldest and most primitive personal relationship, that between buyer and seller, creditor and debtor" (*GM* II: 8). One becomes a debtor when they are under an obligation, leading them to struggle to pay back their indebtedness.

Having explored how different forms of punishment led to the development of conscience, which in turn created bad conscience and raw guilt, Nietzsche revisits his earlier point of view and goes "into the relationship between the present generation and its ancestors" (*GM* II: 19). He suggests that in early tribal societies, people viewed themselves as indebted to their forebears for creating and sustaining their community. They repaid these debts through offerings, celebrations, and adherence to traditional customs. In more severe cases, the repayment occurred through violent and fear-driven methods, including human sacrifice. All of these occurred in fear of the situation that forebearers cease to act "as powerful spirits" and "accord the tribe new advantages and new strength" (Ibid.). Over time, fear was a driving force behind the transformation of venerating ancestral spirits into consecrating mysterious forces, leading to the belief in deities that people fear and obey.

Nietzsche goes on to argue that ancestors became deities, and their descendants continued to show reverence and obedience to them in exchange for their indebtedness. Furthermore, as society grew more powerful and organised, the idea of owing a debt to a deity became ingrained in the culture. Humans inherited the notion from tribe nobility, children from their parents, and even slaves adopted the religions of the deity from their masters. Nonetheless, Nietzsche's assertion about the evolution of the ancestor-tribe relationship is astonishing. According to Nietzsche, "the guilty feeling of indebtedness to the divinity continued to grow for several millennia—always in the same measure as the concept of God and the feeling for divinity increased on earth and was carried to the heights" (*GM* II: 20). The sense of indebtedness reached its peak as faith in God became more prevalent, culminating in the Christian God, who was "the maximum god attained so far" and brought about "the maximum feeling of guilty indebtedness on earth" (Ibid.).

The assertion that the Christian God emerged from conventional deism is startling because it presents an inconsistency within the second essay. Nietzsche explicitly states that the feeling of guilt does not arise from fear of punishment. Punishment is ostensibly intended to elicit a sense of guilt or remorse in the offender, thereby aiding their moral development. In practice, however, punishment is used to intensify fear and subdue humans into behaving peacefully, thereby encouraging individuals to avoid adverse outcomes instead of instilling an awareness of their misbehaviour (*GM* II: 14, 15). If punishing debtors cannot induce a sense of guilt in them, then threatening divine retribution cannot sow guilt in human debtors either. How can this predicament be resolved? If guilt does not stem from fear of punishment by creditor-ancestors, why does Nietzsche assert that guilt has its origin in the creditor-debtor relationship and can be traced back to early ancestor worship?

Although Nietzsche touches on the solution in *GM* II, his discussion about fear of punishment nearly eclipses it. He acknowledges that the concept of God originated from the tribe's fear, but he promptly explains that, as the tribe evolved, God evolved beyond being a mere manifestation of terror, transcending into a personified representation of admirable qualities revered by the tribe. According to Nietzsche, in "the *intermediate* age in which the noble tribes developed—

who indeed paid back their originators, their ancestors (heroes, gods) with interest all the qualities that had become palpable in themselves, the *noble* qualities" (*GM* II: 19). God initially served as a means of coping with the unknown and fearful aspects of existence. People refrained from breaking rules due to the fear of divine punishment, indicating their indebtedness to God. Failing to fulfil one's obligation represented a failure to reimburse the debts owed to God and did not stimulate feelings of self-worth. However, as time wore on, God evolved into a symbol of beneficial human traits representing the tribe's principles and ethical guidelines. Consequently, failing to honour one's obligation would provoke remorse due to a failure to meet the moral standards bestowed and embodied by God.

To clarify this point, an illustration might be useful. For example, when children let down individuals they look up to, for instance, older siblings whom they admire as role models. Usually, children view older siblings who are further along in their development and with whom they share a common language as role models, and imitate their actions subconsciously. Should children fail to fulfil the expectations of their older siblings, causing disappointment, they are not so much afraid of any punishment their older siblings may impose; rather, they will feel remorse and frustrated. One reason may be that errors made by family members normally do not incur severe punishments. The more important reason, though, is that children are more concerned with having let down a role model by acting in a manner inconsistent with their behaviour. Unlike adhering to external norms, following a role model entails a sense of obligation, as people aspire to possess the same traits as those they admire, irrespective of whether there are external incentives to do so. As a study by Katz and Kofkin (1997) shows, children incorporate the values of their role models, which become a significant part of their self-conception. This in turn creates a correlation between obligation and self-esteem, so that the feelings of guilt arising from failing to repay a debt become subjective.

Likewise, neglecting to pay one's debt to a deity who embodies moral principles is akin to failing to live up to the expectations of one's older siblings. The feelings of guilt do not stem from one's indebtedness to more powerful or fearsome deities. Rather, these emotions arise from owing debts to gods who personify and exemplify the standards of personal worth that individuals endorse and aspire to emulate. Since individuals view the obligation to repay their debts as a personal matter, failing to do so results in episodes of bad conscience and internal suffering. The "pang of bad conscience" prompts individuals to hold themselves to expectations that become moral rather than prudential or conventional because they view violating them as "bad", and consider self-punishment to be just and appropriate, based on their perception of themselves as either "good" or "bad." Subsequently, what individuals experience is not solely the "basic affect" of raw guilt but also the "meta-affect" of guilt, which attaches value judgement to raw guilt. By enforcing specific moral expectations on oneself, the moral agent comes into existence.

In summary, guilt requires the judgement of one's inner suffering caused by violation of personal obligations to be merited or deserved. One is indebted to their creditor and is obliged to pay back their debts, just as ancestors created tribes that provided shelters for individuals. When society becomes more powerful, individuals can only turn aggressive instincts inwards, leading to the internalisation of the state of indebtedness and the sense that they deserve punishment for what they did (seek revenge) in the eyes of others. At the later stage, the objective guilt is connected with the remorse of following the role model one admires, which results in the self-imposed judgement upon herself and transforms raw guilt into subjective guilt. Guilt, in this way, includes self-worth as an essential component. It is worth noting that guilt can engage self-worth in two ways. One can feel guilty for what one has done, or one can feel guilty simply for who they are. To judge one's actions to be "bad", one does not have to judge one's character to be bad. This is a crucial distinction because although Nietzsche criticises guilt as an "illness" that "has reached its most terrible and most sublime height" (*GM* II: 19), he also takes the ancient Greek as an example to show that the concept of gods and the feeling of "guilt" can be put to better (or "nobler") use. I will now investigate how ancient Greeks managed "guilt" in a healthier manner, which contrasts sharply with the distorted moralised guilt caused by Christianity.

$$\text{Raw Guilt} \xrightarrow[\text{What One Has Done}]{\text{Indebtedness to gods}} \text{Guilt}$$

**3.3.2 Greek Responsibility**

In Aphorism 23, Nietzsche discusses Greek values and makes a comparison between the Hellenic and Christian ideals to shed light on what "moralised" means. While Nietzsche does not claim that the Greeks experienced guilt in the Christian sense, he describes a value structure in which failure to live up to the standards personified by the gods provoked a powerful emotional response, which functioned as a structural analogue to guilt without entailing the same self-condemning judgement. I call this proto-guilt. Nietzsche believes that its direction was the opposite of Christian guilt. I will provide more arguments for this claim and clarify that the process of indebtedness becoming subjective does not necessarily result in "moralisation". Moreover, taking responsibility for one's behaviour is a healthier attitude compared to Christian religion. Then later, I will investigate how guilt becomes attached to God under the influence of Christianity and shapes humans' indebtedness to God as irreparable.

The main point in Aphorism 23 is that, in ancient Greek society, deities served as a means to alleviate proto-guilt and attain psychological liberation. Ancient Greeks did not equate an individual's failure to express themselves with immorality or wickedness, but rather believed that foolish behaviour was a result of being possessed by a deity. Consequently, attributing any guilt-like emotions to the gods allowed humans to exercise their will more freely.

Some scholars, such as Williams (1993), argue that the Greeks lived in a shame culture rather than a guilt one. From their perspective, this paragraph could be interpreted thus: Greeks did not feel guilt because they attributed bad conscience and the raw guilt resulting from it to the gods. According to Nietzsche, "these Greeks used their gods precisely so as to ward off the 'bad conscience,' so as to be able to rejoice in their freedom of soul" (*GM* II: 23).

However, if the Greeks had not experienced the mechanism that produce accompanying proto-guilt, they would have had no need to attribute blame to the gods. It was precisely because they could feel a form of inner pain or unease that they sought to alleviate suffering by transferring responsibility to divine figures. Some may object that the Greeks projected their misdeeds onto the gods not because they felt guilt-like emotions but because they wanted to exculpate themselves and make blame by others unjustifiable. Yet Nietzsche contradicts this interpretation. According to him, "the gods served as the originators of evil; in those days they took upon themselves, not the punishment but, what is nobler, the guilt" (Ibid.). If the gods did not take the punishment upon themselves, then the Greeks must have done so, but how could one accept punishment without experiencing a sense of proto-guilt and considering oneself responsible for their actions? Therefore, I believe that the Greeks did experience an emotional response analogue to guilt for their actions and took responsibility for them. Nevertheless, their conception of proto-guilt was structurally distinct from the one taught by Christianity.

As previously mentioned, guilt can be associated with one's self-worth based on either their actions or identity. I believe that for Nietzsche, Greeks took responsibility for their actions but refused to accept guilt for who they were. They regarded themselves as having aristocratic descent, belonging to the best society, "being happy, well-constituted, noble, and virtuous" (Ibid.). As reflected in their statement that "how could it actually have happened to heads such as we have" (Ibid.), they rejected the idea that their fallibility made them sinful. Since the Greeks had such high opinions of themselves, they recognised that they might engage in foolish behaviours, but they did not perceive it as an indication of their identity. Therefore, they had no reason to see themselves as evil. They were willing to take responsibility for their actions, and they even admired and encouraged doing so, but solely for what they did.

Strawson's framework provides a more comprehensive explanation of the Greeks' attitudes towards guilt and responsibility. As argued in the last chapter, Strawson draws a distinction between reactive and objective attitudes towards moral responsibility. Reactive attitudes, such as resentment, gratitude, and moral indignation, are emotional responses to the perceived intentions and actions of others. These attitudes are a form of participant attitude, implying that the other person is a member of the moral society and capable of establishing interpersonal relationships. In contrast, objective attitudes involve a more detached, clinical view of individuals as objects of social policy or manipulation. These attitudes may arise because the individual is not an appropriate target for interpersonal relationships and demands for goodwill when she is under abnormal strains, neurotic, or morally undeveloped. In such situations, we observe the person from a distinct standpoint because she is hindered in forming regular interpersonal connections. Applying this framework to the Greeks, we see that while they believed that their wrongful behaviours were attributed to the gods, it was not an excuse that could exempt themselves from responsibility, because it neither indicated that the behaviour was accidental nor that the subject had intellectual defects or an underdeveloped moral sense. The Greeks possessed a mechanism that enabled them to reflect upon their misdeeds and take on punishments and demonstrated reactive attitudes towards their misdeeds, showing remorse and a willingness to atone. However, they did not adopt an objective attitude towards themselves; they did not punish themselves through exile from the community. Instead, they still saw themselves as integral members of the community and strove to repay their debts within interpersonal relationships by undergoing punishment. Hence, the Greeks' approach aimed not to exclude the wrongdoer but to restore social integration through personal atonement. While the Greeks held themselves to moral expectations and took responsibility for their actions, they did not experience guilt for who they were. Alternatively, they would compensate for their indebtedness through punishments, which differed greatly from the concept of 'sin' and 'debt to God' in the Christian tradition. In this way, the Greeks employed guilt-like emotions in a much healthier way.

### 3.4 Moralised Guilt

Moralised and non-moralised guilt follow a similar pattern: God/gods serves as the basis for the internalisation of behavioural norms within the community and eventually becomes the embodiment and representative of these norms. As a result, the members of society strive to emulate and compare themselves to these norms. The difference lies in the fact that in non-moralised guilt, individuals can redeem their debts to gods by undergoing inner suffering and punishment. In contrast, once guilt is moralised, it becomes a permanent state that the individual cannot repay. As Nietzsche argues, moralisation is due to the combination of God and the bad conscience. The concepts of guilt and duty "push back into the conscience... the involvement of the bad conscience with the concept of god" (*GM* II: 21). This means that God becomes the creator of the bad conscience in a moralised sense that generates irredeemable inner suffering, and represents virtues that humans are unable to attain. Since individuals can never possess the same virtues as God, their guilt feelings extend beyond what they have done and expand to encompass their very existence.

### 3.4.1 From Yaweh to God

Nietzsche doesn't expound explicitly in *GM* about how the Greek gods evolved into the Christian God, but in *Antichrist*, he elaborates on how the bad conscience "has reached its most terrible and most sublime height" (*GM* II: 19) and gave rise to the Christian God. Jewish Yaweh serves as an intermediary within this process.

According to Nietzsche, initially, Israelites maintained a natural relationship with all things. Like the Greek gods, Yaweh was a projection of the Israelites' noble qualities:

> Their Yaweh was the expression of their consciousness of power, of their delight in themselves, their hopes of themselves: in him they anticipated victory and salvation, with him they trusted that nature would provide what the people needed – above all rain. Yaweh is the God of Israel and consequently the God of justice: the logic of every nation that is in power and has a good conscience about it. (*A* 25)

However, this period did not last long because the Jews underwent many siege battles throughout their history. If God is the embodiment of a nation's self-confidence and grants unconditional blessings to humans, why would Jewish people still suffer so much? The priests took advantage of this issue and clarified it by interpreting all the Jewish fortune as rewards from God and all their misfortunes as punishments for their disobedience. The priests thus created a new order of the moral world where "natural causality from the world by means of reward and punishment" was banished, so that one "requires an anti-natural causality: all the remaining unnaturalness follows forthwith" (*A* 25). As a consequence,

> every natural custom, every natural institution (state, administration of justice, marriage, tending of the sick and poor), every requirement presented by the instinct for life, in short everything valuable in itself, becomes utterly valueless, inimical to value through the parasitism of the priest (or the 'moral world-order'): a sanction is subsequently required—a value-bestowing power is needed which denies the natural quality in these things and only by doing so is able to create a value… The priest disvalues, dissanctifies nature: it is only at the price of this that he exists at all. – Disobedience of God, that is to say of the priest, of 'the Law', now acquires the name 'sin'; the means of 'becoming reconciled again with God' are, as is only to be expected, means by which subjection to the priest is only more thoroughly guaranteed (*A* 26)

Yaweh turned into a mechanism for evoking the bad conscience in Jewish people. They considered themselves guilty of causing Yaweh displeasure because of their debts arising from their failure to meet the standards for personal worth and virtue prescribed by Judaism's priestly cast. This indebtedness caused inner suffering within the Jewish community. Thus, the Jewish people not only assess themselves objectively based on their wrongdoing, but also view themselves subjectively as guilty because they regard obedience to Yaweh as a personal obligation.

Additionally, the guilt felt by Jewish people isn't limited to the misdeeds they committed but also encompasses their identity. Unlike the Greeks, who ascribed their wrongdoings to the gods and believed in their elevated and virtuous qualities, the Jews' self-censure stemmed from their moral failure, which is why they attached the notion of "sin" to their feelings of guilt. For them, Yahweh chastises them due to their immoral nature, rather than their immoral actions. Jewish morality is no longer just an anthropomorphism of the gods' virtue, but a covenant with the divine.

The process of moralisation did not stop here. Building on Jewish morality, Christianity went on to strip away the last remaining elements of reality to create a more abstract world. Although the Jewish people believed they were guilty, they believed that their relationship with Yaweh was, in principle, expiable and could be repaired by following a new covenant. They viewed themselves as "holy people," and "the chosen people" (*A* 27). In other words, they were the destined and unique ones who were able to become holy like Yaweh and demonstrate his virtues by fulfilling the new law. However, Christianity went further: "as Christianity it negated the last remaining form of reality, the 'holy people', the 'chosen people', the Jewish reality itself" (Ibid.).

Similar to the Jewish God, the idea of the Christian God made people believe that sin was the cause of their suffering. However, Christian guilt goes far beyond Jewish guilt because it represents the belief that there is no such thing as "holy people". For Christianity, human indebtedness to God is unrepayable, because it cultivated "a madness of the will which is absolutely unexampled: the will of man to find himself guilty and reprehensible to a degree that can never be atoned for; his will to think himself punished without any possibility of the punishment becoming equal to the guilt." (*GM* II: 23). The Christian God, as a divine ideal of perfection, is a goal that humans cannot achieve. Therefore, if the Jewish God represents virtue that has not yet been attainable, then the Christian God symbolises the impossibility of achieving perfect and holy virtue. Confronted with such a holy creditor, one had no choice but deeply implant sin into her conscience, which was "within him like a polyp, and then "gives rise to the conception of irredeemable penance", widely known as"eternal punishment" (*GM* II: 21).

$$\text{Raw Guilt} \xrightarrow[\text{Who One Is}]{\text{Indebtedness to God}} \text{Moralised Guilt}$$

As a result, Christianity invented "an even more abstract form of existence, an even more unreal vision of the world than one conditioned by an organised Church." (*A* 27). In this abstract world, it is not human behaviour but the human character itself that is the object of guilt. Since holy perfection is not possible for humans to achieve, they are judged inherently guilty. As Nietzsche states, "the beginning of the human race, its primal ancestor who is from now on burdened with a curse ("Adam," "original sin," "unfreedom of the will"), or of nature from whose womb mankind arose and into whom the principle of evil is projected from now on ("the diabolising of nature")', or of existence in general, which is now considered *worthless as such*" (*GM* II: 23). Therefore, moralisation is not a process in which indebtedness becomes guilt, but rather a process whereby "guilt" and "duty" are combined with self-worth in relation to unattainable human ideals. In other words, it is a process of combining guilt and duty with the conception God, resulting in an inescapable human condition. Moralised guilt is not guilt for what one has done, nor is it only guilt for who one is, but guilt for who one must be.

The comparison between Hellenic and Christian guilt/proto-guilt indicates that Nietzsche offers both a positive and a negative conception of responsibility. Although Nietzsche shows that their ways of taking responsibility differs because of their linking guilt/proto-guilt with self-worth in different ways, he does not provide a sustained analytic differentiation that would reveal the underlying reasons for their divergence, resulting in the situation that his moral attitude—conveyed through poetic and evocative language—remains somewhat ambiguous. It is precisely here that Strawson's framework offers a powerful interpretive tool. Through it, we can uncover that there exist two distinct psychological structures underpinning the forms of moral responsibility Nietzsche respectively praises and condemns. The form of responsibility Nietzsche affirms—exemplified by ancient Greek culture—reflects what we might call a non-pathological framework. In attributing fault to the gods or to external forces, the Greeks shielded themselves from the self-directed aggression characteristic of bad conscience. This did not prevent them from experiencing guilt-like reactive emotions such as resentment or regret; rather, they continued to see themselves as participants in a moral community. Their emotional responses remained intact, but these responses were directed toward their actions rather than toward their very existence. In this sense, the Greeks responded to their wrongdoing with an attitude grounded in their responsibility as a participant in a shared moral life. The Greek interpretative framework allowed for reactive attitudes to function within healthy bounds without devolving into pathological self-condemnation. It thereby preserved the moral affective landscape while avoiding the damaging internalisation that defines the Christian form of guilt Nietzsche so vehemently rejects. By contrast, Christian morality operates within what may be described as a pathological framework of internalisation. It conceives of sin not merely as a moral failing linked to specific actions, but as a condition intrinsic to human nature itself: one is not simply a wrongdoer, but a sinner by essence. This interpretative shift transforms the drive to aggression into an internalised form of hostility, producing what Nietzsche famously calls bad conscience, which is a psychological state resulting in self-contempt, internal reproach, and affective self-denial. In Strawsonian terms, the reactive attitudes here become self-directed, involving a deeply affective mode of self-blame, shame, and guilt—not merely for what one has done, but for who one is. The agent comes to see themselves not as a participant in a moral community responding to shared expectations, but as the object of moral condemnation. This over-internalisation undermines the participant stance and gives rise instead to a quasi-objective attitude directed inward, resulting in a form of alienation from the self. Bad conscience, in this sense, is not a natural response to wrongdoing but a distorted and destructive misuse of reactive attitudes. Strawson's framework thus helps clarify how moral responsibility, far from being a metaphysical property grounded in free will, emerges through affective practices embedded within cultural and interpretative frameworks. In the Greek context, internalised moral emotions remain linked to one's actions and enable the agent to maintain a healthy stance of self-accountability without collapsing into self-condemnation. By contrast, distorted by the Christian religion, internalisation takes the form of self-punishment, whereby the moral failure is experienced as a defect in one's very being. In this way, Strawson offers conceptual tools for understanding Nietzsche's distinction between healthy and pathological forms of responsibility, which are not merely rhetorical exaggerations, but are psychologically and affectively grounded differences in the moral constitution of agency. Responsibility, on this account, is revealed to be not simply a metaphysical burden, but a product of the interaction between social practices and psychological attitudes.

## 4. Conclusion: Nietzsche and Strawson

As mentioned at the beginning of this dissertation, traditional debates on moral responsibility often focus on the conditions necessary for moral responsibility to be compatible with determinism. This search for conceptual conditions can turn into an endless intellectual game, since pessimists can always find a counterexample that meets the compatibilism-friendly condition for moral agents yet still appears incompatible with determinism. This is where the Strawsonian approach stands out. Instead of focusing on conceptual analysis, Strawson bases moral responsibility on everyday experiences, specifically on our reactive attitudes towards others' actions and underlying intentions. This perspective sheds light on Nietzsche's views on moral responsibility. As previously argued, Nietzsche denies moral responsibility only within the context of Christian morality, not all moral responsibility. Through a Strawsonian lens, we can see more clearly how Nietzsche situates the process of moral formation within interpersonal interactions and reveals why Christian morality, as Nietzsche describes it, represses life.

More specifically, while Strawson bases moral responsibility on our reactive attitudes towards others' actions, Nietzsche bases it on the reactive attitude towards one's own actions: the feeling of guilt. For Strawson, our reactive emotions towards intentions behind actions indicate one's moral responsibility. Nietzsche similarly links reactive attitudes with moral responsibility: when one's actions fall short of self-expectations, bad conscience contributes to the feeling of guilt for one's actions and sometimes even self-worth as a person. This process gives rise to both the healthier guilt like the one the Greeks took upon themselves and the Christian moralised guilt criticised by Nietzsche, indicating the moral responsibility one bears for what one has done or who one is.

Moreover, both Nietzsche and Strawson emphasise that morality arises from interpersonal relationships, whether these relationships are between individuals or within oneself. For Strawson, we adopt a participant attitude towards those we consider to be members of our moral community. Similarly, Nietzsche describes how the Greeks saw themselves as members of the moral community, believed in their capacity to meet moral expectations, and thus, even when they made mistakes, could repay their debts and restore social relationships through punishment and personal effort. However, while Strawson suggests that we may modify or even suspend reactive attitudes towards those outside the moral community, Nietzsche argues that expelling oneself from the moral community marks the peak of moralised guilt. At this point, individuals believe that their guilt is inherent and irredeemable, and their personal relationship with the moral community can no longer be repaired, giving space to the priests to create a more abstract vision of the world. In this way, Nietzsche distinguishes two affective orientations: one in which failure invokes personal responsibility within a framework of self-respect, and another in which failure implicates the whole self in a system of moral condemnation. While only the latter corresponds to what Nietzsche calls "moralisation", both illustrate the emergence of moral agency, since in both cases individuals view the obligation to repay their debts as a personal matter and their failure to do so results in episodes of bad conscience, attaching value judgement to their material indebtedness. If this reading is correct, then what underlies moral agency, according to Nietzsche, is our distinctive capacity to experience suffering in the form of self-evaluation—more specifically, the sense that one *ought* to suffer for having breached a commitment. Without such a sensibility, the very possibility of moral agency would collapse. In other words, it is our affective responsiveness to material indebtedness as internalised commitment that grounds our capacity for responsibility.

With these considerations in mind, we can better understand how Nietzsche takes a Strawsonian reversal. Both Strawson and Nietzsche shift our attention away from conceptual constructions toward moral practices. Specifically, Strawson examines the reactive attitudes we adopt in interpersonal relationships in response to others' actions and intentions, and how these responses shift under certain conditions, in order to justify moral responsibility. Nietzsche, by contrast, investigates moral practices in the form of punishment and their corresponding reactive affects, namely guilt, as a way of exploring the moral content of responsibility. Neither thinker seeks the preconditions that make moral agency; instead, both prioritise the explanatory order of holding responsible, contending that it is the practice of holding responsible that reveals the origin and structure of morality. It must be emphasised, however, that Strawson and Nietzsche adopt significantly different methodologies. Strawson situates moral responsibility synchronically within ordinary interpersonal relationships, while Nietzsche places it diachronically within historical development. This divergence reflects their differing aims: Strawson's project is defensive, aiming to justify our practices of responsibility in the face of the challenge posed by determinism. He embeds responsibility in everyday moral interactions to demonstrate its practical character and thereby sidestep metaphysical debates. Nietzsche's project, by contrast, is critical—he challenges our habitual understanding of moral responsibility by tracing its historical origins and the ideological function of guilt. His central question is not whether responsibility is justified, but why we have come to experience it as meaningful in the first place. Despite their differences in direction and intent, Strawson's framework provides a valuable tool for understanding Nietzsche's critique. Rather than softening Nietzsche's attack on modern morality, the Strawsonian lens helps to clarify the psychological structure underlying different moral forms within Nietzsche's genealogy, which helps to illuminate the implicit content of "being responsible" in various contexts.

Nietzsche's description of the gradual development of morality also aids Strawson in addressing the issue of appropriateness. As mentioned in the last chapter, Strawson faces the question of whether children should be held responsible if moral responsibility is closely related to reactive attitudes. Would our reactive attitudes, as emotions, always be reliable? How do we determine who is an appropriate target for moral attribution? An effective approach is the Borderline Theory, offering borderline-style answers to questions such as "should very young children be appropriate targets of moral blame?" According to Strawson, children are "creatures that are potentially and increasingly capable both of holding, and being objects of, the full range of human and moral attitudes, but are not yet truly capable of either" (Strawson, 2008: 32). As a result, the treatment of children must comprise a kind of compromise, situated between normal reactive attitudes and the objective attitude. Strawson's account focuses on the development of individual morality, tracing its progression from an immature to a mature stage, while Nietzsche examines how morality evolves from its early beginnings to more advanced forms throughout human history. Just as individuals grow from childhood, where they merely follow rules without experiencing subjective guilt, to adulthood, where personal actions are tied to self-worth and produce genuine feelings of guilt, human history reflects a similar trajectory. This progression marks the maturity of individuals as fully responsible moral agents. Whereas Strawson's essay does not provide a systematic articulation of his borderline theory, Nietzsche's *GM* II can be understood as a systematic expression and complementary perspective of this theory. The internalisation of conscience constitutes the dividing line. Before internalisation, humans recognise their wrongs but view themselves from a detached, objective standpoint, unable to see themselves as fully responsible moral agents. Through internalisation, individuals, for the first time, come to truly understand their responsibilities and expectations for themselves, making them fully capable of bearing moral responsibility. Thus, we can understand how one becomes an appropriate target for moral attribution through historical development without relying on conditions external to human agency.

In summary, Strawson supplies a descriptive vocabulary (reactive attitudes, participant/objective stance) that clarifies *how* responsibility functions as an affective practice, whereas Nietzsche supplies the diachronic, genealogical account of *why* those affective structures emerged; together they illuminate the psychological mechanics of responsibility. Nietzsche's description of moral development in *GM* II not only renders the threat of determinism to moral responsibility irrelevant, as determinism remains constant over time and thus cannot explain the evolving conditions that determine the appropriateness of holding someone responsible, but also complements and vividly illustrates Strawson's borderline theory. Furthermore, by reinterpreting moral responsibility through a Strawsonian lens, we gain a clearer understanding of Nietzsche's view of moral responsibility as a practice rather than a conceptual construct. Thus, we can know that Nietzsche does not straightforwardly reject all forms of morality, nor does he see responsibility as inherently burdensome in every context. For Nietzsche, moral responsibility can only achieve a healthy state through practice, enabling individuals to take pride in bearing responsibility for their actions. It is also through practice that moral responsibility encounters the possibility of becoming moralised, which reveals why Nietzsche both affirms and critiques morality. The kind of moral responsibility he endorses lays the groundwork for the emergence of souls capable of possessing free will.

# Chapter IV

## Value, Valuing and Value Creation

### 1. Introduction

In the previous chapter, I investigated Nietzsche's description of the historical evolution of moral responsibility in *On the Genealogy of Morals*, from the primitive stages of conscience to the complex sense of guilt in Christian morality, through a Strawsonian lens. As shown in the development process, the ancient Greeks started to associate guilt with their self-worth by taking responsibility for their actions. They were proud enough to repay their misdeeds and reconstruct their relationship with society. Now we grasp the kind of moral responsibility Nietzsche supports. However, merely advocating for the healthy moral responsibility of the ancient Greeks does not fully support Nietzsche's concept of free will. The sovereign individual described by Nietzsche is not only capable of bearing responsibility, as the Greeks did, but does so on the foundation of self-created values. Thus, while the ancient Greeks exhibited the courage to take responsibility, Nietzsche does not simply advocate for a return to their past ideals. Instead, he praises those who look forward, such as the sovereign individual, who possesses the strength to create their own measure of values (*GM* II: 2). This forward-looking attitude necessitates an exploration of the conditions under which such creative power emerges, fostering the ability to transcend traditional social constraints and create new moral frameworks. In the next chapter, when I delve into Nietzsche's notion of free will, I will discuss how the key factors in the development of moral responsibility contribute to value creation and how free will can be understood as a strong will on this basis. However, we must not hastily embrace the concept of value creation simply due to Nietzsche's positive tone, as doing so risks reducing our analysis to mere rhetorical flourish. Therefore, in this chapter, I will argue for the very possibility of value creation from Nietzsche's perspective. Only by doing so can we establish a genuine connection between Nietzsche's conception of moral responsibility and his notion of free will. To investigate value creation, I will first take a look at contemporary views on Nietzschean metaethics. Then I will argue that Nietzsche occupies a position between realism and anti-realism. He maintains that values are subjective and individual, but he also calls for an evaluation of values. The tension between these two metaethical poles captures the significance of Nietzsche's concept of valuing, which gives rise to its own evaluative standard. On this basis, I will investigate how value creation is possible and what value creation is for Nietzsche. I argue that Nietzsche's ideal notion of value creation is an experiment that contributes to the further development of values. Initially, I will examine the value creation by priests in *On the Genealogy of Morals*. I will then argue that this mode of value creation is not the only one. Nietzsche's task of value creation is not predominantly intellectual, but concerns practice under evaluative standards, a process more common than is typically assumed. Finally, I will investigate the relationship between solitude and value creation, arguing that Nietzsche's value creation is group work that requires the involvement of others.

## 2. Beyond the Greeks: Nietzsche's Vision for the Future

As I discussed in the previous chapter, examining the history of moral development through a Strawsonian lens, particularly in relation to conscience and guilt as reactive emotions, reveals Nietzsche's vision of a more robust conception of moral responsibility, which he links to ancient Greeks. This approach also exposes that the morality Nietzsche critiques emerges as a product of intensified moralisation, shaped by its fusion with the concept of God. However, Nietzsche repeatedly emphasises that he is not advocating a return to the past. For instance, in the second essay of *Untimely Meditations*, he writes:

> "We need history, certainly, but we need it for reasons different from those for which the idler in the garden of knowledge needs it, even though he may look nobly down on our rough and charmless needs and requirements. We need it, that is to say, for the sake of life and action, not so as to turn comfortably away from life and action, let alone for the purpose of extenuating the self-seeking life and the base and cowardly action" (*UM* II, Foreword).

For Nietzsche, history, even that which is worthy of admiration, should not be an end in itself. Rather, history should serve life and future action, not nostalgia. This perspective directly opposes any notion of resurrecting past ideals, including those of ancient Greece. However, Nietzsche also contends that individuals are stifled by the poisonous influence of Christian morality. If he rejects the idea of returning to the past in search of an ideal figure embodying healthy moral responsibility as a model, then the question arises: how can individuals possessing free will emerge? When Nietzsche describes what he terms the *sovereign individual*, who is the "master of free will," he asserts that such an individual "possesses his measure of value" (*GM* II: 2) This implies that the individual is not bound by pre-existing value frameworks, whether inherited positively or negatively, but instead establishes their own principles of valuation and actively generates new values accordingly.

Moreover, the sovereign individual is not the sole figure Nietzsche envisions as exemplifying free will. He introduces several other figures, such as the "genuine philosopher". As Nietzsche notes, "The education of a genuine philosopher is a task that demands something different—it demands that he create values" (*BGE* 211). Another key figure is the "overman", perhaps one of Nietzsche's most renowned concepts. Nietzsche famously states, "Man is a rope, tied between beast and overman—a rope over an abyss..."  (*Z*, Prologue, §4). For Nietzsche, the overman does not signify a return to the past with a healthier morality; rather, it represents a forward-oriented transformation. We are situated in a transitional state, detached from humanity's original condition yet profoundly moralised. Yet, this remaining state is precisely what Nietzsche critiques and seeks to liberate us from. Just as the sovereign individual is "the ripest fruit" of the "tremendous process" of moralisation and "morality of custom" (*GM* II: 2), so too is the overman an entity both emerging from and transcending the present. Thus, Nietzsche declares, "The overman may live some day" (*Z*, Prologue, §4), where "some day" inherently signifies a future moment.

A more explicit example is found in Nietzsche's notion of the "philosophers of the future" in *Beyond Good and Evil*. The subtitle "Prelude to a Philosophy of the Future" (Vorspiel einer Philosophie der Zukunft) suggests that the entire book appears to be oriented toward the future, heralding the emergence of a new kind of philosophy and a new type of philosopher—one that, according to Nietzsche, has not truly existed since the times of Heraclitus, Plato, and Empedocles (*BGE* 204). Nietzsche repeatedly writes about these figures, stating, "I see such new philosophers coming up" (*BGE* 2). He also refers to these "philosophers of the future" (*BGE* 42–44) as "*very* free spirits", or calling them "new" philosophers, who are described as "commanders and legislators" (*BGE* 211). All these figures, who transcend present constraints and embody free will, according to Nietzsche, represent a vision for the future rather than a mere revival of the past. Similarly, while Nietzsche praises both, the fundamental distinction between the ancient Greeks and these future figures lies in their ability to create their own values. As Nietzsche writes: "With a creative hand they reach for the future, and all that is and has been becomes a means for them, an instrument, a hammer. Their 'knowing' is creating, their creating is a legislation, their will to truth is—will to power" (Ibid.). In this way, the creation of new value establishes a connection between Nietzsche's affirmative conception of moral responsibility and his idea of free will. Mere taking proud responsibility for one's actions is inadequate; one must assume responsibility in accordance with self-created values to genuinely actualise oneself as an individual with free will.

## 3. Nietzschean Values: Between Realism and Anti-realism

Although Nietzsche advocates for the creation of new values, how exactly can new values be created? The creation of new values raises a metaethical issue: the traditional division between value realism and anti-realism. Realism is the view that value exists in a mind-independent manner. Anti-realism is the denial of this thesis, holding either that (1) values do not exist at all, (2) values do exist but are created by human subjects. However, the focus on metaethics raises more questions than it solves and risks overlooking the significance of Nietzschean values. On the one hand, Nietzsche seems to be an anti-realist about values (*GS* 301). On the other hand, he holds that certain tastes can enjoy privileged status, for he refers to the taste of those he admires as noble, while those displayed by those he does not admire are bad. The puzzle is: If Nietzsche is an anti-realist about value, then there are no human-independent standards to apply. Given the apparent fact that values are merely human perspectives, which have internal good without being grounded externally, Nietzsche's praise of noble values and the concept of value creation seem to lack an evaluative foundation. Yet if Nietzsche is a realist, the key role value creation plays would be minimised because humans cannot create values, but merely discover them. So, does this inconsistency truly exist? Is value creation for Nietzsche merely an uplifting rhetoric rather than a possible practice? To resolve this issue, I propose focusing on the very activity of valuing, rather than conceptually situating Nietzsche within the traditional value realist/anti-realist division. In my view, the practice of valuing itself reveals a possibility: that Nietzsche exists in a space between realism and anti-realism. It is precisely through the process of valuing that evaluative standards emerge, thus reconciling the creation of value with the endorsement of a particular kind of value. To understand how the traditional value realist/anti-realist division fails and how this failure opens up the possibility for an intermediary path, we must first revisit the traditional discussions to examine the tensions created by their failure.

### 3.1 Anti-realism

#### 3.1.1 Hussain's Fictionalism

Many hold that Nietzsche's value realism commits him to an error theory, which means that beliefs expressed in moral judgements are errors because they aim at truth, yet there are no evaluative moral facts to support them. However, attributing an error theory to Nietzsche is problematic because it claims that all values are falsely represented as objective, which undermines Nietzsche's writings on value and value creation. Nadeem Hussain's solution is to combine error theory with fictionalism. According to Hussain, Nietzsche regards all claims of value as false. Hussain argues that several of Nietzsche's passages suggest that nothing in the world has inherent value, and that the only source of value is human projection. For example, in *Twilight of the Idols*, Nietzsche writes that "there are absolutely no moral facts. What moral and religious judgements have in common is the belief in things that are not real. Morality is just an interpretation of certain phenomena or (more accurately) a *misinterpretation*" (*TI* VII: 1). Therefore, Hussain concludes that Nietzsche casts doubt on evaluations of value since all seemingly real values are actually generated by humans.

As mentioned above, there is a tension between error theory and the significance of Nietzsche's concept of value and value creation. Hussain's solution is to interpret Nietzsche's valuing as a "fictionalist simulacrum of valuing" (Hussain, 2007: 158). By simulacrum, Hussain means a pretense or make-believe. As he claims, "Nietzsche's free spirits pretend to value something by regarding it as valuable in itself while knowing that in fact it is not valuable in itself" (Hussain, 2007: 170). For example, when we appreciate an artwork such as a painting, even if we know that it is nothing more than materials organised in a certain way, we do not regard it as valueless; instead, we seriously look for meanings in the art.

However, Hussain's account is unsatisfactory. One problem is that Nietzsche's denial of value appears to be weaker than Hussain suggests. In *TI* which I cited above, Nietzsche regards morality as interpretation. What he seems to claim is not that there are no values about morality at all, but rather that present moral values have been misinterpreted. As can be seen from the Greeks, they had an interpretation of moral values that Nietzsche appreciates. Moreover, if Nietzsche considers something as a misinterpretation, then there must be evaluative standards that define what makes an interpretation mistaken. In this passage, what is required is an evaluative criterion and more accurate moral interpretations, not a denial of the existence of moral values altogether.

Another problem is that fictionalism does not satisfy the demands Nietzsche places on his ideal figures. Nietzsche praises figures such as the sovereign individual and emphasises their ability to create values in order to build an ideal ethical position and construct a response to nihilism. Typical nihilism includes the assumptions that all values are baseless and that there is no meaning in human existence. To counter nihilism, Nietzsche's strategy is to create values that make life worthwhile and meaningful, thereby becoming "human beings who are new, unique, incomparable, who give themselves laws, who create themselves" (*GS* 335) that can transcend the Christian illusory moral framework. However, mere fiction is inadequate as a motivational role. When one knows values to be illusions, she is unlikely to be motivated to act, let alone create a meaningful life. In creating new values, Nietzsche requires more substantive values and admires certain historical figures who embody these values, such as Goethe. Recall that Hussain views valuing as pretence or make-believe. It would be odd if Nietzsche's ideal figures ignored what things truly are and went around pretending.

### 3.1.2 Leiter's Naturalism

Because fictionalism is not plausible, many interpret Nietzsche as a naturalist and connect his naturalism to anti-realism. Brian Leiter offers an influential reading of this sort and provides two arguments to support his view. One is the "inference to the best explanation", and the other is based on type-facts. I will investigate these two arguments in turn.

#### 3.1.2.1 Inference to the Best Explanation

Leiter's inference to the best explanation argument is based on the history of intractable disagreement in moral philosophy. The skeleton structure of this argument is as follows:

(1) If there were objective moral values, epistemically well-situated observers would eventually converge upon them after two thousand years;
(2) Contemporary philosophers, as the beneficiaries of two thousand years of philosophy, are epistemically well-situated observers;
(3) Contemporary philosophers have not converged upon moral values;
(4) Conclusion: there are no objective moral values (IBE from the first three premises).

However, based on both philosophical reasoning and of Nietzsche's texts, I am not convinced that moral disagreement entails anti-realism. Firstly, Leiter does not give a precise explanation as to when his "eventually" might occur? The lack of convergence so far does not imply that convergence cannot be reached at all. When confronted with a difficult question, philosophers may be on the way to finding answers, and the process will take a significant amount of time. By contrast, convergence is not equivalent to the existence of objective values. For example, most philosophers have converged upon the claim that knowledge is justified true belief. However, Gettier's essay (Gettier, 1963) proves that this claim many had converged upon is not objectively true.

In addition, I do not think Nietzsche's discussion of the multiple views held by philosophers is concerned with moral disagreement. Leiter (2010) cites passages in *Beyond Good and Evil*, especially section 5, 6, 186 and 187, to support his view.

> They all pose as if they had discovered and reached their real opinions through the self-development of a cold, pure, divinely unconcerned dialectic…while at bottom it is an assumption, a hunch, a kind of "inspiration"—most often a desire of the heart that has been filtered and made abstract—that they defend with reasons sought after the fact. They are all advocates who don't want to be called by that name, and for the most part even wily spokesman for their prejudices which they baptize "truths." (*BGE* 5)

Leiter argues that Nietzsche regards this disagreement as located at the heart of the positions of most Western moral philosophers. This intractable disagreement highlights the foundational difference between moral philosophy and science. He also observes that the "rational foundation for morality" is "merely a scholarly variation of the common *faith* in the prevalent morality" (*BGE* 186). These passages address moral disagreement, but I believe Nietzsche's purpose in these passages is to illuminate the relationship between moral values and philosophers. Philosophers, indeed, have not converged on moral values. However, their aim is not to search for moral values at all. What motivates their philosophical endeavours is not a "drive to knowledge", but rather another drive that has "employed understanding (and misunderstanding) as a tool" (*BGE* 6) to create the world they desire. They are not "honest enough in their work" when they claim to address "the problem of truthfulness" (*BGE* 5). Philosophy, in this sense, becomes "a personal confession of its author" (*BGE* 6).

Therefore, Nietzsche does not critique the significance of philosophy or deny the possibility of reaching convergence on moral values. Rather, the problem lies in the character of philosophers, who treat philosophy merely as a tool and fail to honestly pursue the moral values they believe to be true. Consequently, it becomes impossible to reach any consensus (regardless of whether objective moral values truly exist). Therefore, Nietzsche aims to cultivate a new breed of philosophers—philosophers of the future—who are more self-motivated to discover themselves and more creative in forging their own values. The challenge for Leiter lies in demonstrating how to move from Nietzsche's critique of philosophers to the denial of all moral values.

#### 3.1.2.2 Type-facts

Although Leiter's inference to the best explanation argument aims to prove a lack of objective moral facts, he does not think that this lack will result in a loss of the ability to explain moral experience. According to Leiter, Nietzsche holds that higher humans are harmed by morality in the pejorative sense because morality is a barrier to their flourishing. Therefore, his aim is to free higher humans from the shackles of pejorative morality. Given the anti-realism position that Leiter attributes to Nietzsche, the evaluative standard to distinguish higher from lower humans is not an objective principle but results from "doctrine of types"  (Leiter, 2002: 91). Leiter calls these constitutions "natural" facts or "type-facts". Recall Leiter's view on Nietzsche's fatalism discussed in the first chapter. I explained that type-facts are "facts about the person's unconscious drives or affects" (ibid.), which fix a person's life trajectory. Because of the stability of these type-facts, the psycho-physical constitution of each person has been fixed from the start, "which defines him as a particular type of person" (ibid.). In this view, Leiter holds that Nietzsche's notion of higher humans is not that they embody objective higher values, but that their psycho-physical constitution and life trajectory are distinctive. Nietzsche can distinguish higher from lower without being committed to a universal moral standard because all values are determined by type-facts.

Some may oppose the position that although, according to Leiter, Nietzsche views our basic moral responses as brute artefacts of our psychological constitution, there is nothing in Nietzsche's view to rule out the possibility that more complicated moral feelings could involve a cognitive component. Moral judgements arise not only from a basic affect, which is defined as "inclination towards or aversions" regarding certain states or actions, but from the interaction of the basic affect and affective response to basic affect—the meta-affect that may include a cognitive component. Leiter himself acknowledges this claim, but he thinks that this will not pose a threat to seeing Nietzsche as a naturalist. He cites passages from Nietzsche's text to support the view that Nietzsche's failure to be a thorough-going non-cognitivist about the affects under moral values can explain why a revision of our beliefs may lead to a revision of our feelings. "We have to learn to think differently [umzulernen]—in order at least, perhaps very late on, to attain even more: to feel differently [umzufühlen]" (*D* 103). For Leiter, the affects now have two different sources. There are affects rooted in physiological facts and there are those inculcated by local culture and beliefs, but these cognitive components may also be "dependent on affective investment", meaning that moral judgements "are *still* symptoms of the affects" (Leiter, 2013).

However, type-facts are not as stable as Leiter suggests and may not be sufficient to determine a person's life trajectory. Nietzsche seems to encourage individuals to access knowledge and "make it instinctive" (*GS* 11), implying that instincts themselves can be modified. By integrating new knowledge into our instincts, original instincts are transformed, suggesting that the physiological attributes represented by type-facts are also subject to change. In *GM*, Nietzsche also describes the emergence of bad conscience as the redirection of cruel instincts in response to social demands, which indicates that a person's instinctual nature and life trajectory are more fluid than Leiter's interpretation allows. If changes in environment and culture do not result in changes of life trajectories and humans remain the same when time passes, there seems to be no meaning for Nietzsche's investigation of how humans became who they are through his genealogical method.

In addition, Leiter refers to Jonathan Haidt's empirical research on the "social intuitionist" model of moral judgement to argue that reasons do not explain moral judgement. According to Haidt's research (2001), moral judgements are produced by emotional or affective responses in most ordinary situations. Other empirical research by Fiery Cushman, Liane Young, and Joshua D. Greene (2010) also suggests that when people have "selective deficits in emotional processing" due to brain disease or injury, they tend to make different moral judgements about hypothetical situations, like the famous trolley problems[6], compared to those with typical emotional responses.

[6] *The trolley problem* is a well-known thought experiment in ethics. It involves a scenario where a runaway trolley is headed towards five people on the railtracks. There also exists the choice to divert the trolley onto another track, where only one person is tied up. The dilemma raises questions about whether it is more ethical to actively cause harm to save more lives or to avoid taking action, even if it means more people will be hurt.

This finding lends support to Leiter's view that moral values are symptoms and sign-languages of affects. However, I find the conclusions of these experiments unconvincing, as it is possible that moral judgements are influenced by, but not solely determined by, affects. Furthermore, reasoning may also directly impact moral judgements. Studies on cognitive load demonstrate that when reasoning is compromised—such as when individuals are distracted or overloaded—they are more likely to rely on emotional responses rather than reasoned analysis, resulting in different moral decisions (Greene et al., 2008). This indicates that deficits in reasoning, alongside emotional processing, may contribute to the differences in moral judgements, contradicting Leiter's view.

#### 3.1.2.3 Conclusion on Leiter's Reading

Leiter has provided a seemingly promising account of Nietzsche's naturalism and anti-realism. He proposes the inference to the best explanation argument to claim that moral values are best described in terms of psychological facts without appealing to unnecessary or over-complicated objective values. He also insists that values depend on immutable type-facts, which provide the ground for distinguishing higher from lower humans. Against Leiter, I have argued that type-facts are not immutable as Leiter claims. Leiter's inference to the best explanation argument is also unconvincing on both philosophical and textual grounds. Based on these points, I contend that Leiter's views are not only unpromising, but also fail to grasp Nietzsche's focus on value creation.

### 3.2 Realism

I have argued that contemporary anti-realist interpretations of Nietzsche on moral values, including fictionalism and naturalism, result in more problems than they solve. However, I will argue that this need not lead to the conclusion that Nietzsche subscribes to realism. Attributing objectivity to values seems to undermine the role of value creation, which is an essential aspect of

Nietzsche's philosophy. A closer examination of objective interpretations of Nietzsche may reveal the complexity of Nietzsche's value theory.

Despite Nietzsche's advocacy for pluralism and multiple perspectives, his critique of morality, particularly Christian morality, suggests that he upholds some standards of good and bad or right and wrong. Some propose the idea of substantive realism, arguing that Nietzsche uses a single value standard to justify all other values, such as power, life, and freedom. Others known as subjective realists, maintain that Nietzsche'c critiques serve as internal standards he uses to demonstrate how certain values fail to meet their ideal demands.

### 3.2.1 Substantive Realism

Tamsin Shaw (2007: 78-136) offers a promising account of this view. She argues that, for Nietzsche, the primary concern is the overall development of humanity. The criticism Nietzsche provides concerning moral values is in this sense universal. In this regard, Shaw suggests that for Nietzsche, there exists a fact about the evaluation of values, which is oriented around a single universalist commitment: that of the flourishing, and ultimately self-overcoming, of humanity. However, she also maintains that Nietzsche distinguishes between human sub-types and does not believe that everyone is capable of arriving at real moral values. This leads to a problem: how can Nietzsche's universalism be reconciled with the pluralism of its basic orientation?

To address this issue, Shaw categorises Nietzsche's view on values as an example of what T. M. Scanlon refers to as parametric universalism (Scanlon, 1998: 328-329). According to Scanlon, there are ultimate standards that also allow for context-dependent variations to determine what is right. Unlike relativism, it emphasises "applying a fixed set of substantive moral principles to varying circumstances" (Ibid.). For instance, helping others is an ultimate standard for moral judgement. However, "failing to help a person whose car has broken down... would be a serious wrong in a place where someone who is stranded overnight is likely to freeze to death, but not a serious wrong in a safe country with a mild climate" (Ibid.).

Turning to Nietzsche, Shaw's strategy is first to argue that Nietzsche's work centres on the ambition of "the enhancement of the species "man" (*BGE* 44). Among the factors that can lead to the degeneration of the human type, the lack of differentiation plays a crucial role. He warns that "every enhancement of the type 'man' has so far been the work of an aristocratic society—and it will be so again and again—a society that believes in the long ladder of an order of rank and differences in value between man and man" (*BGE* 257). Since uniformity can only be achieved by levelling down high types, he opposes any moral or political views such as egalitarianism and democracy. Instead, the best way to achieve this goal is through differentiation and multiple perspectives. Although Nietzsche contends that values such as ascetic morality may be beneficial for some but disastrously harmful for others, his point is not that certain values are inherently bad, but that human flourishing can best be achieved when some people value mastery and strength while others value subordination and weakness. As he claims, "The essential thing in a good and healthy aristocracy is that it... accepts with a good conscience the sacrifice of untold human beings who, for its sake, must be reduced and lowered to incomplete human beings, to slaves, to instruments" (*BGE*, 258). Aristocratic society, in turn, contributes to human flourishing. To support the claim that this value judgement has universal validity and is not merely applicable to certain individuals, Shaw argues that Nietzsche presupposes the existence of intrinsic values within humanity—values that inherently rank some people as superior and others as subordinate. Just as a star shines regardless of whether anyone observes it, Nietzsche implies that greatness exists independently of human perception (*BGE*, 285). This suggests that the evaluative judgement Nietzsche proposes is substantively real, and the process of life-enhancement and liberation from the anti-realistic illusions of religion lies in uncovering and affirming the true hierarchical order of human values.

Shaw's approach, relating universalism with pluralism through parametric universalism, is a brilliant insight. However, I do not believe this category can be attributed to Nietzsche. For Nietzsche, though some individuals can determine what good values are for themselves, others are constrained by their intellectual capacities, leading them to be trapped in illusions sustained by religion and social conformity. Thus, it is difficult to imagine how "fettered spirits" (*HTH* I: 225) pursuing weak values they deem appropriate for themselves could contribute to the enhancement of humanity as a whole. Recall Nietzsche's notorious comments on women, in which he suggests that women should embrace weakness and be "kept under control by their fear of men" (*BGE* 221). Even if it were the case that obeying men were the best choice for women at a certain stage of history, I see no way that their obedience would lead to their own flourishing any more than to the flourishing of all humans, in the long run.

### 3.2.2 Subjective/Internal Realism

Similar to substantive realists, subjective realists also support the existence of evaluation standards. What is different is that they regard these standards as internal to the values we already possess. John Richardson advocates this kind of internal realism and argues that Nietzsche occupies a position between extreme internalism and externalism.

How can Nietzsche reconcile his internalism with his externalism? How can the subjectivity of Nietzsche's values be reconciled with his rankings of values? Richardson's strategy begins with interpreting Nietzsche as a neo-Darwinist who identifies three main ways of valuing, which are differentiated by their types of selection. These are the body's value, the moral agent's value and one's own value. We can also call them animal, human and superhuman values (Richardson, 2004).

Richardson argues that body values are the most basic values selected through the evolutionary process. As Nietzsche writes, body values manifest in our "muscles, nerves, ganglia" (*WP* 373; *D* 119). According to Richardson, valuing is an intrinsic part of aiming or will. The ultimate goal of will is growth; that is, will aims to extend its power to control other drives that related to valuation. All animals, he contends, share the same fundamental aim of growth and are naturally selected to respond to outcomes that enhance their power. In other words, to achieve their aim of growth, they instinctively value activities that more effectively serve the underlying will to power. Drives, defined as dispositions toward specific actions, are distinguished by their particular activities or projects that carry values of their own, but all drives share the common value of the will to power. For example, an amoeba aims at power through its responsive disposition to consume nearby organisms as food, while trees aims at power by enjoying sunlight. They value different activities, but for both the will to power is the fundamental end guiding all valuation.

Richardson further argues that human values, in contrast to animal values, are signs of consciousness through words. They result not from instincts but from conscious social practices like sharing and imitating. He claims that the symbolic feature of human values serves to promote agential control. To achieve this aim, "they wish the faculty that answers to them to be dominant in the person, to rule the competing aims in the drives" (Richardson, 2020: 21). However, despite promoting agency, human values are shaped primarily for the benefit of the community rather than the individual. The great work of "taming," as Richardson describes it, has shaped values in favour of group interests. These values require individuals to be collaborative members of a group and function as signals indicating the rules they must obey in order to serve the group. This, he suggests, creates an internal conflict within humans, as animal values promote individual preservation, whereas human values prioritise group interests. As a result, humans become distanced from their bodies. According to Richardson, Nietzsche believes that this deep-rooted opposition to bodily drives has led human values to take on a predominantly ascetic character, one that fosters dangerous impulses and ultimately contributes to the sickness Nietzsche sees in humanity.

Moreover, Richardson contends that human values may not always accurately reflect what truly serves the interests of their deepest drive. For example, slaves mistakenly take ascetic values as beneficial to their well-being. Therefore, one of Nietzsche's aims, as Richardson reconstructs it, is to overcome the conflict between human values and bodily drives by incorporating truths about values into the act of valuing. Nietzsche approaches this problem, according to Richardson, by positing superhuman values, which are meta-values that govern the evaluation of one's first-order values. Meta-values themselves are more than Nietzsche's "valueds" and enjoy a special status that allows them to justify other values, as these are values that we already have.

Among these meta-values, Richardson identifies power as one of the most fundamental. As mentioned above, power, the will to grow, functions as life's ultimate meta-value, ensuring that values are aligned with truth. Another essential meta-value, he argues, is the pursuit of truth itself. These meta-values, such as power and truth, are embedded within all other values and serve to refine them. By recognising that we deeply value power and truth, Richardson suggests, we will be able to apply this criterion not only to our judgements of body values but also to our evaluation of human values, thereby bringing the latter into better alignment with the former. In this sense, valuing becomes an active process that enhances values to better serve the demands of power and truth, as reflected in all animals' responses to outcomes that can enhance their power. This interpretation from Richardson combines elements of both internalism and externalism. The externalist aspect emerges from the evaluative standards we use to compare our new values with old ones, which reflects an evolutionary improvement in our values—new values are selected for their ability to better achieve power and truth. However, while these new values are better positioned to achieve power and truth, Richardson does not take them to be objective in the traditional sense. Rather, their validity stems from the act of valuing itself, based on their effectiveness in advancing power and truth, rather than being rooted in any objective or external standard.

There are several problems with Richardson's account. First, he regards values merely as directions for what one aims at and as the end of drives. However, in the following section, I will argue that values play a more positive role in shaping one's drives, rather than simply setting their aims. Another issue is that I find the process of value creation through evolutionary self-selection inadequate as a response to the problem of nihilism. Nihilism is more than "valuelessness" (Richardson, 2004: 124); it is fundamentally about meaninglessness, where "the aim is lacking" and "'why?' finds no answer" (*WP* 2). What is required for addressing nihilism is not merely a unifying project but something that can inspire love and commitment. I will discuss this point in the final chapter when addressing free will.

**3.3 Conclusion on Realist Readings**

It is challenging for anti-realism to grasp the significance Nietzsche attributes to values and the act of valuing. For fictionalism and naturalism, values are mere errors of projections devoid of evaluative standards, making the seriousness of value creation hollow. For realism, the values inherent in value creation are rarely grasped, and values are often linked to individuality. Therefore, to categorise Nietzsche's view within either of these two metaethics positions loses sight of key aspects of his philosophy. In the subsequent sections, I will position Nietzsche's values within the middle-ground between realism and anti-realism,  identifying value creation as the defining feature of valuing.

## 4. Value and Valuing

### 4.1 The relationship between Taste and Value

The German word "Geschmack" or its English translation "taste" is often used in a descriptive sense, referring to the objective qualities of things like food or drink. For example, when discussing the taste of a dish, we typically describe its inherent qualities (salt, sour, sweet…), without necessarily implying personal preference (good or bad)[7]. Yet, Nietzsche frequently uses 'taste' to refer to personal preferences or judgments, where it entails an evaluation—something absent in the former case, which is merely a neutral description. A clear example of this evaluative usage can be found in Pierre Bourdieu's work, *Distinction: A Social Critique of the Judgement of Taste,* where he perceives "taste" as an expression of social distinctions and hierarchies in an evaluative sense, rather than a purely neutral description.

By showing how valuing overlaps with tastes, Nietzsche elevates taste to a legitimate source of justification. As he puts it, taste can "count as an argument" (*HAH* II: 183). Opinions, along with their proofs and refutations, are all symptoms of tastes (*GS* 39). One interpretative claim motivated

[7] "The taste is good" is an evaluative judgement, but taste here refers to an objective quality, describing the flavour of food.

by such passages is that Nietzsche regards many evaluative judgements as expressive of a particular mode of taste—that is, of a distinctive way of valuing. In this sense, taste functions as a microcosm of one's value orientation—for Nietzsche, an instance of a particular taste is synonymous with an evaluative perspective.

The relationship between valuing and tastes is more clearly illustrated in this passage concerning Nietzsche's reaction to the Bible:

> In the Jewish "Old Testament," the book of divine justice, there are human beings, things, and speeches in so grand a style that Greek and Indian literature have nothing to compare with it. With terror and reverence one stands before these tremendous remnants of what man once was, and will have sad thoughts about ancient Asia and its protruding little peninsula Europe, which wants by all means to signify as against Asia the "progress of man." To be sure, whoever is himself merely a meager, tame domestic animal and knows only the needs of domestic animals (like our educated people of today, including the Christians of "educated" Christianity) has no cause for amazement or sorrow among these ruins—the taste for the Old Testament is a touchstone for "great" and "small"—perhaps he will find the New Testament, the book of grace, still rather more after his heart (it contains a lot of the real, tender, musty true-believer and small-soul smell). To have glued this New Testament, a kind of rococo of taste in every respect, to the Old Testament to make one book, as the "Bible," as "the book par excellence"— that is perhaps the greatest audacity and "sin against the spirit" that literary Europe has on its conscience. (*BGE* 52)

People like Nietzsche react to the Old Testament with "terror and reverence" and find it "more after his heart". Modern individuals, by contrast, have no reason to experience the same reactions when encountering the Old Testament. As Nietzsche suggests, differing tastes for the same thing reveal the interplay between one's evaluative judgements and underlying mode of valuing. Thus, taste is not a superficial preference but a concrete expression of the same normative structures that also underpin valuing. It is not fitting for individuals with "small" mode of valuing to react to the Old Testament in the same way as those with "great" valuing. Their tastes are not false, but simply unsuited to individuals like them, who are more inclined toward "a kind of rococo of taste". The evaluative judgements one makes (to appreciate one type but not others) are indicative of one's fundamental valuing, which are expressed through one's tastes. Thus, differences of taste illustrate not only what values one affirms but also how one stands in relation to them: whether in a grand style or in narrower mode. Taste thereby provides a concrete analogy for Nietzsche's later distinction between good and bad valuing.

From a practical perspective, taste may be a better starting point for my argument because it is more accessible. While values and valuing often occupy a more abstract status, taste is something we frequently encounter in our daily lives. To explain what I mean by "taste," I can hand you a spoon and let you try a dish; to explain a "value", or the mode of valuing, I cannot simply put it on a plate. In this sense, taste is a more familiar and concrete phenomenon than "value" in its abstract form. Unlike valuing, which often operates through abstract conventions and lacks tangible presence, taste is something we can directly experience—both sensorially and evaluatively—by perceiving and then appraising. Even when taste carries evaluative elements, its operation can be grasped in a concrete act: tasting a dish and finding it harmonious, or hearing a melody and finding it graceful. This familiarity and accessibility with the concept of taste makes it easier to understand and relate to. Moreover, taste and valuing share the same structural features: both guide selection, preference, and evaluation, and both, as I will argue later, involve an internal standard that can distinguish between better and worse instances. Yet taste is not only structurally similar to valuing but can also be regarded as its lower-order, experientially accessible manifestation—the form in which these evaluative structures first become palpable. While taste is rooted in individual sensibility, it already carries a sense of structure and direction, shaping how we evaluate and respond. Clarifying taste's structure, therefore, will serve as a stepping stone to clarifying the more abstract structure of valuing. Therefore, I intend to explore and maintain the tension between realism and anti-realism by beginning with taste, arguing that while taste is deeply rooted in everyday experience and can be judged by personal preference, it also possesses a degree of internal structure and normativity that goes beyond mere subjective preference. Once this is established for taste, I will transpose the same logic to valuing, and show that valuing too can be assessed by an internal standard, which will later be articulated as the constitutive aim of valuing itself. To be more specific, the standards for judging valuing are intrinsic to valuing itself, rather than existing as external, objective criteria, which allows valuing to be assessed and refined from within, without appealing to any external, objective criteria.

### 4.1.1 Taste

Nietzsche often uses taste to refer to physiological processes in a gustatory sense (*D* 188). As mentioned above, he more frequently refers to taste as an expression of individual preferences and associates it with evaluations. In many such cases, taste is a visceral reaction—an instinctual response that manifests as love or disgust without the need for reflective judgement[8]. For example, recall Nietzsche's texts on the Old Testament, he describes how the "great" individuals react to it with an immediate sense of awe and reverence—this reaction is a form of taste that arises spontaneously, without the need for contemplation. In contrast, those who are "tame" and "domesticated" respond to the New Testament with a similarly immediate, but less profound reaction of preference without reflection. In this sense, at its lowest register, taste is merely physiological, reflecting fundamental instincts and drives.

Similarly, Leiter takes a Humean reading that views tastes as Humean sentiments. The evaluative contents of tastes, according to this view, pertain to one's type-facts at the personal level. However, I do not think that tastes are nothing more than expressions of individual drives and desires, which would render tastes entirely subjective and beyond evaluation. While Nietzsche indeed considers tastes to be outcomes of drives and desires, he also points out that tastes can develop beyond the limit of physiological instincts. As I explored in more detail in the last chapter, "all instincts that do not discharge themselves outwardly turn inward" (*GM* II: 11), where the taste for cruelty driven by natural instincts was transformed into the first appearance of the human soul. Instincts represent the most basic stage of taste (Geschmack in seiner untersten Stufe), implying that tastes can develop into more complex and nuanced forms (NF-1880, 11, [164]).

To see how tastes can develop into a more complicated form, it would be better to consider an example of taste in art. Wandering in the National Gallery, a person may be unreflectively attracted by the colourisations or vigour of Turner's oil paintings. She may have no pre-knowledge

[8] By using "reflective judgement", I do not mean it in the Kantian sense where taste is one of its expressions. Reflective judgement here just refers to a type of judgement that incorporates the cognitive component of reflection.

of the history of art, but the expressive brushwork full of strength and imagination attracts her eyes and arouses admiration in her. Her taste in Turner's paintings may lead her to prefer Romantic painting. After this visit, she may read more related books, visit more galleries and gain a more conscious appreciation for this school. This taste might also develop beyond Romantic painting and lead to new tastes in Expressionism, not only in painting but also in music, literature and architecture. Over time, her current tastes seem to have little connection with her initial taste in Turner. Yet, these new tastes are developed from her initial taste, which is the projection of physiological factors and can be traced back to desires and drives. What is different is that they are qualitatively more complicated: as more faculties and forms of consciousness enter into play, taste undergoes a transformation. What began as a rudimentary projection of drives gradually becomes a reflective orientation. In other words, tastes are projections of drives and desires but cannot be reduced to them. What emerges here is not a mere string of likings but a nascent evaluative standpoint: the subject begins to *rank*, generalise, and transfer standards across cases. In other words, taste is *valuing in embryo*: an affectively mediated, increasingly reflective orientation that already does the work values do.

Building on this idea, Peter Poellner offers an influential cognitivist reading on taste. He claims that it is appropriate to interpret taste for Nietzsche as a "pattern of conscious affectivity" (Poellner, 2007: 231), where taste is constituted by affects (Affekte), or more precisely, by certain recurrent affective evaluations that form what we can call a general sensibility. One with a certain taste will typically evaluate objects of the corresponding type as either positive or negative on the basis of evaluative features derived from recurrent conscious experiences of affects. Taste as a pattern of conscious affectivity is not merely a collection of feelings but rather evaluations based on conscious affective experiences. This does not only concern taste. When Nietzsche asks what the definition of evaluation is, he gives the answer that moral evaluation is an interpretation that is caused by physiological conditions as well as certain mental judgements. But who are interpreters? The answer is: Affects! As he writes, "moral evaluation is an *exegesis,* a way of interpreting. The exegesis itself is a symptom of certain physiological conditions, likewise of a particular spiritual level of prevalent judgments: Who interprets?—Our affects" (*WP* 254). In Nietzsche's own words, affect plays a central role in tastes and evaluation and infuses taste with a broader dimension of conscious awareness. On this picture, taste just is the phenomenal mode in which valuing first appear to us: a structured pattern of favouring/disfavouring, yet already exceeds mere preference by introducing a nascent orientation that guides ranking, selection, and commitment..

To understand taste as Poellner describes it, it is essential to clarify what Nietzsche means by "affect." For Poellner, although Nietzsche uses the term "affects" mainly in the sense of emotions, he also retains "broad usage of 'affect' for any mental episode which constitutively involves a pro- or con-attitude (or as I shall say, a favouring or disfavouring) with a distinctive phenomenology" (Poellner, 2007: 229). Specifically, he argues that affects are not merely instinctual reactions but provide a bridge between mere feeling and the assessment of value. Affects are seen as emotional responses that are also evaluative—they penetrate into the process of evaluative judgement and help us determine what we find valuable or distasteful. As he writes, "Nietzsche expresses the idea that our basic mode of acquaintance with value is through 'affects' (Affekte) in many places. In a minimal interpretation, this point is of course implied by his general perspectivism, according to which any conceptualisation, including any classification of items in evaluative terms, essentially involves affectivity" (Ibid.). By understanding affects in this sense, Nietzschean tastes move beyond pure physiological factors and transforms into more complex constructs that are built upon emotional foundations and involve conscious evaluative processes.

Evidence in support of this interpretation can be seen in several passages. For example, Nietzsche claims that tastes for the Christian religion partly consists in the feeling of fear, and partly in reactive affects. He writes that "Das Jenseits wird mit Bildern dekorirt, die theils die Schöpfungen der unmittelbaren Furcht, theils der reaktiven Affekte sind" (NF-1875, 9[1])[9]. He also claims that taste as a response and sensation reveals something deeper, which is "the whole oder of rank of its values" in one's soul (*BGE* 268). As Nietzsche says, "Whoever wants a demonstration of the 'German soul' ad oculos should merely look into German taste" (*BGE* 244). Given these passages, it seems promising to interpret Nietzschean tastes as a complex interplay between conscious evaluations and deeper, instinctual forces. When Nietzsche says taste reveals "the whole order of rank of one's values" (*BGE* 268), he is not adding the element of valuing to taste from the outside; he is identifying taste as the primary site where valuing becomes manifest and assessable.

However, one issue with Poellner's interpretation is that regarding taste as conscious affectivity seems to lead to an anti-realism perspective, ultimately concluding that nothing provides Nietzschean taste with privileged status in contrast to other epistemic standards. While I do not fully agree with Poellner's understanding of taste, I agree with him that affects in Nietzsche's texts should not be equated with pure sentiments, as doing so would shift the argument from "tastes are based on affects" to anti-realism. Accordingly, if taste is already a mode of valuing, then assessing taste is not extra-aesthetic moralising but the critical articulation of its *internal standards*—the features that make taste better or worse as taste itself. To address Poellner's issue while ensuring that taste maintains epistemic privileging, in §4.1.2, I will explore the evaluative standards of taste

[9] The reason for using German here is that there is no translation in English of this passage.

that Nietzsche proposes, which I think are internal, and provide textual evidence to support this reading of the assessment of taste.

### 4.1.2 The Assessment of Taste

I think some passages cited by the anti-realists do not, upon further examination, support their views. For example, Leiter cites a passage in *Thus Spoke Zarathustra* that pertains to disputes over taste. Nietzsche says that taste has something "to be disputed" and is 'at the same time weight and scales and weigher' (*Z*, "On Those Who are Sublime", 116). As mentioned above, Leiter takes dispute as the premise for the best argument in favour of anti-realism. According to Leiter, Nietzsche's response to different evaluative matters should be to 'turn his back and say, "Oh well – doesn't share my evaluative tastes"' (Leiter, 2015: 124). However, if this is the case, then the 'dispute' itself seems to lack significance. In this passage, Zarathustra's emphasis on disputes is a response to those who think that taste is "nothing to be disputed". He may praise dispute because he believes that disputes set evaluation standards along which judgments can be made. One can consistently hold that she is right in her values while others are wrong, according to epistemic lines, and try to persuade them. Yet, even if Leiter's account is problematic, this passage does not provide direct evidence for the reading that taste possesses epistemic privilege.

One piece of direct textual evidence casting doubt on reading Nietzschean affects as pure sentiments is in the third essay of *GM*. Here, Nietzsche writes that "perspectives and affective interpretations" can be employed  "in the service of knowledge". He also emphasises that although we do not possess objective knowing but merely perspectival seeing, "the more affects we allow to speak about one thing, the more eyes, different eyes; we can use to observe one thing, the more complete will our 'concept' of this thing, our 'objctivity,' be" (*GM* III: 12). If affects are nothing more than sentiments devoid of cognitive elements and incapable of epistemic differentiation, then the claim in this passage would be difficult to reconcile with the anti-realism reading. Affects in this sense can contribute to human understanding in a way merely subjective emotions cannot.

Taking affects as mere sentiments is not Nietzsche's opinion. By contrast, he directly equates "taste and tact" with "reality" in *The Anti-Christ*. And "good and delicate taste and tact" seems to be connected with "the whole integrity of knowledge", which "we have won back for ourselves" against Christian morality (*A* 59). In other passages, he describes figures he admires such as Goethe, "the last German before whom I feel reverence" (*TI* VIII: 51), as also possessing a sense of reality (*TI* VIII: 49). Nietzsche praises Goethe for his "attempt to overcome the eighteenth century through a return to nature" (Ibid.), and the aim of this self-overcoming is to strive for something "strong" (Ibid.) that pertains to "greatness" (*TI* VIII: 50), namely, "reverence for everything factual" (Ibid.). Along with this reading, one may have good reason to think that there exists an evaluative standard to differentiate strong from weak and that the noble tastes Goethe possesses enjoy some kind of epistemic privilege.

Then what is the assessment of taste? Where does the assessment come from? In a letter to Heinrich Köselitz, Nietzsche acknowledges that he has "no reason, no logic, no imperative" for his taste for Köselitz's music. Yet I do not think that his claim can eliminate the existence of evaluative principles. Given Nietzsche's urgent will to understand this problem[10], he is certainly not suggesting that taste is merely a matter of personal preference and that there is no possible explanation of one's taste. His inability to explain his appreciation for a certain kind of music indicates the possibility

[10] As Nietzsche writes in his letter,"Diese Tage gingen Sie mir sehr durch den Kopf: ich hätte Sie gerne da gehabt, um Aesthetica mit Ihnen zu reden". eKGWB: Letter to Heinrich Köselitz (Peter Gast), 19 November 1886.

that the standard by which he judges aesthetic taste is not something that can be objectively grasped or fully explained through reason alone, but appeals to a more fundamental level of evaluative reasons, probably intrinsic to affective-evaluative features themselves. In other words, the standard by which Nietzsche evaluates music may be something that is accessible only through the internal, subjective experience of taste itself, much like something that can only be felt or intuited, not fully captured in words or rational analysis.

I find this thought promising, but what counts as "intrinsic" is far from clear. One possible line of thought is that, given that affects are not merely simple emotions but can also be moved up to reflective feelings that involve evaluative dimensions, one might introduce the concept of meta-affects to address this problem. Similar to meta-values that evaluate one's first-order values, meta-affects are affects that respond to first-order affects, representing a higher level of reflective feelings that help to frame a criterion for judgements. To see if there exist different responses to the same affective emotion, consider the following passage:

> The same drive evolves into the painful feeling of cowardice under the impress of the reproach custom has imposed upon this drive; or into the pleasant feeling of humility if it happens that a custom such as the Christian has taken it to its heart and called it good. That is to say, it is attended by either a good or a bad conscience! In it itself it has, like every drive, neither this moral character nor any moral character at all, nor even a definite attendant sensation of pleasure or displeasure: it acquires all this, as its second nature, only when it enters into relations with drives already baptized good or evil or is noted as a quality of beings the people has already evaluated and determined in a moral sense [ . . . ] (*D* 38)

Here Nietzsche claims that the same affect has the potential to give rise to two totally different affective results. The first-order affect in this situation may be something like "to avoid offending enemies". If experienced by a Greek, he will judge his response as expressing cowardice and feels ashamed. By contrast, a Christian would regard it as a moral virtue of humility. There are two layers of affects here: first, there is the simple affect of aversion towards offending enemies, which is direct reflection of drives and desires, but then there is the distinctive moral affect of feeling ashamed (as the Homeric Greek does) or proud (as the Christian does) of that instinctive affective response. The central claim of this approach is that our meta-affective responses can provide us with the "intrinsic phenomenal (dis)value of that state" (Poellner, 2007, 247) when those first-order states that we can characterise as the attractiveness or unattractiveness are represented to ourselves.

But what counts as the evaluative matter in this approach? Jonathan Mitchell (2017) contends that the only probable source depends on whether the first-order affect, the experience of attractiveness or unattractiveness, is accurately represented to a conscious subject. Accurate is interpreted as "without self-deception". Mitchell proposes an example of egoism. Egoism, he claims, tends to distort the evaluative principle for self-interest. For instance, confronted with a smarter person who constitutes a threat, the egoist would feel angry and provide a self-justification of their anger. She will claim that the defective emotion is justified, but pretend to be a good person. Yet, if the egoist were to accurately represent both the intentional object and the corresponding affective response to herself, she would experience her jealousy as unattractive and judge her intrinsic phenomenon, her aversion to this person, as not valuable.

As Mitchell puts it, Nietzsche also proposes this kind of evaluative principle. He supports his idea through his interpretation of *BGE* 263.

> There is an instinct for rank, that, more than anything else, is itself the sign of high rank; there is a pleasure in nuances of respect that indicates a noble origin and noble habits. The subtlety, quality, and stature of a soul is put dangerously to the test when something of the first rank passes by … Anyone whose task and exercise is the investigation of souls will use this very art, in a variety of forms, to establish the ultimate value of a soul, the unalterable, inborn order of rank it belongs to: this sort of investigator will test out the soul's instinct. (*BGE* 263, see also *BGE* 260, 265, 163)

Mitchell re-phrases "pleasure" as "meta-affect" and "nuances of respect" as "first-order affect", re-interpreting this passage as follows: "there is an affective attunement to intrinsic phenomenological value ('instinct for rank') that is itself the sign of an individual being of high rank" (Mitchell, 2016: 27). In this account, Mitchell contends that Nietzsche takes the attunement to phenomenological value as the evaluative principle to test one's "soul's instinct", to see if she takes an affective attunement to the instinct for rank when her affective response to the object of high value is accurately represented to her. In other words, the Homeric Greek displays the requisite ability, the attunement to phenomenological value, by feeling shame and regarding the aversion to offending enemies as cowardice.

I take this approach to be similar to Harry Frankfurt's view on free will. Frankfurt distinguishes between first-order desires and second-order desires. First-order desires have actions as their objects, while second-order desires, known as desires about desires, are concerned with the desires of first-order desires. As Frankfurt puts it, one acts freely if and only if her second-order desires identify with first-order desires. Although "attunement to phenomenological value" provides an explanation to "identify with", I think that the meta-affect approach faces the same problems as Frankfurt's theory. Even if we can define what "accurate" means regarding the representation of first-order affects, how can we establish if our affective responses to first-order affects are correct in the sense of forming "attunement to phenomenological value"? Is it necessary to propose a third-order affect in order to solve this problem? Then what about fourth-order, fifth-orderr, and so on? Furthermore, this approach sounds as though a noble person prefers nobility, and a weak person prefers weakness. But this presupposes the difference between strong and weak, which makes the approach circular.

In this regard, I argue that instead of construing Nietzsche's concern in terms of "meta-affects," we should turn to his account of *taste* itself. For Nietzsche, taste gradually develops into an intrinsic framework of evaluation, insofar as it begins to organise and disciplines our drives from within, so that new orientations and standards of appraisal can emerge. One of its central features is the drive toward unity, but unity here should be seen less as an end in itself than as a condition for further creation. A passage from *GS* 290 shows that Nietzsche regards "giving style to one's character" as a great art governed and formed by "a single taste" (*BGE* 290). Nathan Drapela notes that in the German edition, where 'Ein Geschmack' is translated as 'a single taste,' Nietzsche capitalises on the article "Ein", which means "one" in English. He does so probably to emphasise the singularity of taste and to indicate that the key to success is the governing of one taste. One taste takes the lead to form an overall coherence of one's intrinsic dispositions, and this coherence is precisely what enables the cultivation of further tastes and gives rise to one's style. Far from being a mere restriction of possibilities, the unified taste provides the organisational strength through which new perspectives and evaluative orientations can be generated. In this sense, to "give style to one's character" is not simply to impose order, but to initiate new forms of evaluation and to reorganise the normative structures that guide one's sensibility. Style emerges when coherence itself becomes productive, enabling the creation of further tastes rather than merely fixing what already exists. The characterisation of taste as a self-imposed principle of unification also clarifies Nietzsche's concern about the disputes in tastes. Recall that Nietzsche takes tastes as having something "to be disputed" and is "at the same time weight and scales and weigher" (*Z*, "On Those Who are Sublime", 116). The morphological relation between these three words suggests that this is a dispute internal to taste itself but not between diverse tastes: taste is at the same time agent, object, and means.

In this view, as long as one's experience of the world is conditioned by one's affective constitutions, there is no need to look for an external standard to which one can compare oneself. To assess whether one's affective responses, such as tastes, are appropriate, one must determine whether they can foster development and give "style to one's character" (*GS* 290) by operating as a unifying whole that integrates into one's affective constitution. Humans are born with a set of affective properties that shape their preferences and aversions. This makeup of an individual's affective structure defines the kind of person one is. However, these affective dispositions are inherited from various sources and exist in a disorganised form. To give style to one's character and develop one's character requires prioritising certain affective dispositions, thereby creating coherence out of an otherwise unstructured mass. Since this coherence is an ongoing process of integration, which makes possible further transformation, it follows that character is not statically determined by type-facts, as Leiter suggests. Rather, the development of one's taste and evaluative judgement is directed toward a dynamic integration that enables further growth. The intrinsic evaluative principle of taste is therefore not exhausted by unification; rather, unification functions as the condition under which new refinements, perspectives, and evaluative orientations can emerge.

I will elaborate later on what exactly unity is and how it facilitates the further development of taste. For now, the essential idea is that the evaluative principle may be intrinsic to taste itself, constitutive of its capacity for further development and refinement. By refinement here I do not mean a merely decorative polishing of existing tastes, but the generative process through which unity enables new, more coherent possibilities to emerge. Such generation is already a refinement insofar as it enhances and integrates one's evaluative orientation. Internal standards open the possibility of drawing a line between appropriate tastes and those out of bounds. Based on this idea, I argue that the intrinsic standard by which tastes can be evaluated lies in its capacity for development and refinement through unity. Good tastes are those that achieve coherence among previously disorganised affective tendencies in such a way that this unity itself becomes a basis for further refinement and new perspectives. In this sense, unity is the dynamic condition that makes the further development of taste possible.

To see this, consider the case of aesthetic disagreement. Imagine you are discussing a painting with your friend. You argue that Van Gogh's *Starry Night* is the greatest painting ever, while your friend prefers Vermeer's *The Little Street*. There are many ways you can find to persuade your friend. You could point to Van Gogh's immense influence—not only on Expressionism but on modern art as a whole. Additionally, you could discuss his use of colour, the organisation of the scene, and the imaginative quality of his brushwork. Although it is difficult to reach a definitive agreement, this discussion itself provides an opportunity for both of you to deepen your appreciation and understanding of painting as an art form. To continue this discussion, however, presupposes that you already share a set of standards for making aesthetic evaluations and judgments. You are engaging within a common framework of affective coherence, meaning that both of you possess the fundamental ability to appreciate art from an aesthetic perspective. Your disagreement stems not from a lack of coherence in the standards of art appreciation but from differences in taste and perspective. Within this framework, disagreement becomes productive, since it refines sensibilities and expands interpretive horizons. By contrast, if the debate shifted to popularity—arguing, for instance, that Van Gogh is more famous worldwide—then the evaluative discussion would collapse. What disappears is not merely coherence but the generative function of taste: its capacity to integrate experience in a way that opens new perspectives on art appreciation. Without this, the disagreement loses its significance and devolves into a meaningless comparison.

The example of the painting illustrates that although you and your friend cannot agree on which painting is better, it is still conceivable for a reasonable person to prefer either one. However, to extend this example further, imagine that another friend, Sam, comes to join you and insists that Van Gogh's work is no match for Giacometti's sculptures, such as *Walking Man*. Even if Sam provides reasons to justify his preference, you are unlikely to regard his opinion as a valid contribution to the discussion. Sam violates the implicit standards of this discussion because he introduces sculpture into a conversation about painting, an art form governed by different aesthetic criteria. By introducing sculpture into a debate about painting, he shifts the evaluative ground and undermines the conditions that allow the discussion and disagreement to generate refinement and deeper appreciation. What collapses here is not only coherence but the very function of taste—its power to organise experience in ways that yield perspectives capable of further refinement and deeper integration. Without such conditions, the exchange of ideas ceases to foster development—

where "development" means the cultivation of more coherent and discriminating judgements, or a richer and more penetrating appreciation—and degenerates into a meaningless clash of preferences.

The agreed-upon standards intrinsic to the debate itself set the boundaries for disputes over taste. They rule out irrelevant or disruptive answers, such as appeals to popularity or shifts to a different art form, while still allowing space for genuine disagreement. Within these boundaries, intrinsic standards do not necessarily provide a precise criterion for determining what constitutes good taste. To ask whether a preference for Van Gogh's *Starry Night* is objectively better than a preference for Vermeer's *The Little Street* is to misconstrue the nature of the dispute. Nietzsche's point can be read differently: the evaluative principles of taste are not universal standards that rank tastes in terms of their correctness but conditions internal to the activity of judging itself. In other words, there are constitutive features of aesthetic engagement that make the practice of taste possible at all, and these features provide the ground on which good and bad tastes can be distinguished, not by appeal to an external measure but by how far a given taste contributes to the internal standards. The discussion of paintings from a shared aesthetic perspective fosters refinement and generative expansion of sensibility in art, while those that violate it disrupt the discussion, reducing it to a quarrel that undermines its purpose.

Therefore, I hold that good tastes are those that cohere into a unified structure capable of sustaining and expanding the possibility of further tastes. Bad tastes, by contrast, undermine this generative process and may even corrupt the very context in which new tastes could arise. This is not to say that bad tastes are not tastes at all; rather, they are forms of taste that obstruct refinement and undermine the cultivation of possibilities. As Nietzsche writes in *Thus Spoke Zarathustra*, what he admires are "choosy tongues and stomachs". For him, "those who consider everything good" have no sensitivity to the world, "but chew and digest everything". Indiscriminate consumption is not the absence of taste, but a form of taste that disables refinement. It is not fair to say that a swine has no taste, but by chewing food in "the swine's manner" (*Z*, "On the Spirit of Gravity", 194), it is unlikely to develop discrimination and appreciation. The analysis of taste thus clarifies how an activity can be governed from within by standards that are constitutive of its further development. Given the structural parallels between the activity of taste and valuing, and the fact that taste, as a practice, represents the more elementary form of valuing, this insight carries over to valuing. Valuing, too, is assessed not by appeal to external criteria, but by its own internal standard—what I will later articulate as its constitutive aim. Building on this, I will then turn to two limiting conditions of valuing.

**4.2 Back to Value**

As argued above, a particular taste is synonymous with an evaluative perspective. I also contend that tastes can develop beyond physiological factors into a more complex form, suggesting that evaluative perspectives also consist of elements other than drives. Moreover, I have also argued that taste furnishes an intrinsic standard of evaluation: it disciplines and organises drives into a unified structure. Such unity does not guarantee novelty, but it provides the necessary condition for new orientations to emerge, since coherence is what allows further reorganisation and transformation. Since valuing and taste share this structure, the same applies to valuing. Valuing, while rooted in psycho-physical facts, cannot be reduced to the activity of these facts and thereby open up the possibility of becoming increasingly complex over time, as physiological drives are progressively organised, reflected upon, and integrated with conscious evaluative activity. They are not merely mechanistic by-products of drives, nor simply goal-setting devices, but active perspectives, which is an "Optik", through which experience is interpreted, coloured, and reorganised. In *Anti-Christ*, Hollingdale's translation of the German word "Optik" as "perspective" captures the significance Nietzsche attributes to perspective. As Nietzsche writes, "Out of this erroneous perspective [Optik] on all things one makes a morality, a virtue, a holiness for oneself, one unites the good conscience with seeing falsely—one demands that no other kind of perspective [Optik] shall be accorded any value after one has rendered one's own sacrosanct with the names 'God', 'redemption', 'eternity'" (*AC* 9). By Optic, Nietzsche means a framework through which one engages in evaluations. These frameworks not only shape perception but also open the possibility of generating further values. In this sense, valuing, like tastes, can be assessed internally: good valuing are those that sustain the activity of valuation, while bad valuing are hostile to the very activity of valuation. Nietzsche's discussion of Christian morality can be read as a critique of a mode of valuing that brings about "a hostile separation from all spiritual height" (*GM* III: 4) and is even "hostile to life" (*GM* II: 24). It is not wrong because it fails to match an external standard, but because it corrodes the very conditions for the possibility of revaluation.

Since the activity of taste contains an internal standard by which it can be assessed, and valuing and taste share a similar structure, I argue that the same applies to valuing: the standard for their assessment is internal to the activity of valuing itself. This brings us to the central problem: if valuing itself provides the standards for its assessment, how can such standards ground normative claims about valuing without collapsing into objectivism? To clarify the whole picture of value evaluation, I propose to develop a Nietzschean form of constitutivism.

### 4.2.1 Constitutivism

Certain kinds of activities are distinguished from other activities in that participants of these activities necessarily have particular aims. For example, trying to win the game is the aim of what constitutes the essence of a basketball game. All other aspects of basketball, such as scoring, defending, rebounding and assisting contribute to this aim, but it is not sufficient to count as playing basketball that one just scores, defends, rebounds and assists according to the rules of basketball. In other words, the possibility of being a basketball game cannot exist without the pursuit of victory. If you are merely dribbling or shooting for fun without aiming to win, then you are not really engaging in a basketball game. Therefore, winning the game is a constitutive aim of playing a basketball game, which is an inescapable element that defines the activity as what it is. It is true that sometimes the constitutive aim may be influenced by other factors. For example, if you are teaching a child who just starts learning the game, you may adopt the aim of letting the child experience the fun of that game, which will definitely modify the way in which you pursue the aim of winning. You may deliberately miss shots or refrain from defending with full effort so as to give the child a better chance of winning. But you are still making an effort, even just at a minimal level, to win the game. You will still make shots and score, you will still defend against the opponent's offense, and in this way you demonstrate to the child what a basketball game is. If you never approach the aim of winning at all, then you are not engaged in playing basketball, but rather in the activity of teaching the child how to play basketball. In this sense, the aim of winning is non-optional for basketball players: if you are playing basketball, then you have this aim. On this basis, a constitutive aim can be defined as follows:

> (Constitutive Aim) Let A be a type of attitude or event. Let G be a goal. Then A constitutively aims at G iff
>
> (i) each token of A aims at G, and
>
> (ii) aiming at G is part of what constitutes an attitude or event as a token of A. (Katsafanas, 2011: 629)

Let us now examine the relationship between aims and evaluative standards. It is widely accepted that aims generate evaluative standards of success. In other words, we can evaluate A with regard to whether they achieve G. However, this claim applies to all aims, in addition to constitutive aims. To give an example, if the aim of playing a basketball game is to win the game, it is reasonable to evaluate the game based on how well players achieve the goal of winning. Suppose, in another situation, one not only has the aim of winning the game, but also has the aim of learning some new moves or simply enjoying the process, then there are several standards of success: it is reasonable to evaluate certain moves with regard to whether these moves bring one closer to winning, whether they enable one to acquire the skills necessary to play the game well, or whether they make the experience of playing enjoyable. What is special about constitutive aims is that they are intrinsic to the event or activity in question. In other words, they are non-optional for this activity. It is conceivable to play basketball games without aiming to learn something new or merely enjoy it, but you cannot play basketball without aiming to win, even just at a minimal level. Failure to aim at winning undermines the event or activity as what it is. Thus, constitutive aims generate intrinsic evaluative standard of success. Given that winning is an inescapable and non-optional constitutive aim of basketball, it follows that the reasons derived from this aim are likewise inescapable and, in this sense, intrinsic to the activity itself. This principle extends beyond basketball games and applies to activities in general. Constitutive aims define what it means to engage in activities, which thereby generate intrinsic standards of success.

This raises a crucial question: how does the intrinsic nature of evaluative standards contribute to resolving the paradox between value realism and anti-realism? More specifically, if Nietzsche, as he repeatedly asserts, denies the existence of objective values (*GS* 301, *D* 210, *TI* VII: 1, *BGE* 108), then all the values humans praise are merely individual perspectives. This creates tension with the privileged status accorded to certain modes of valuing and reduces Nietzsche's commendation of noble values to mere personal preference. Conversely, if Nietzsche adopts a realist stance, values would only be discovered rather than created. Thus, the puzzle emerges: how to reconcile Nietzsche's central claims about value? If all values are human creations, how can some values nevertheless count as better than others?

Katsafanas summarises the three claims as follows:

(1) Power has a privileged normative status.

(2) There are no objective values, or there are no objective facts about what is valuable.

(3) All values are created by human activities.

(Katsafanas, 2013: 152)

Let us assume, for the sake of argument, that a particular mode of valuing A holds a privileged status for Nietzsche:

(1) Valuing A, as well as the value it creates, has a privileged normative status.[11]

(2) There are no objective values, or there are no objective facts about what is valuable.

(3) All values are the products of human valuing activities.

[11] Nietzsche does distinguish between higher and lower values, but this distinction is not grounded in the objective content of the values themselves. What matters is the mode of valuing. Thus, it is the quality of valuing that provides the standard, not the content of value taken in isolation. Good valuing produce good values, but this does not confer objectivity upon the values themselves.

Suppose one aims to identify the constitutive aim of valuing. The first step is to demonstrate that a specific aim, A, is present in every instance of valuing. Moreover, just as winning constitutes playing a basketball game, a certain aim, A, is necessary to valuing and determines its very possibility, defining it as such. In other words, there is no valuing without the aim A. Aim A constitutes the essence that every episode of valuing embodies. Since the constitutive aim generates the internal standard of success, the evaluative assessment of valuing can be derived from aim A. On this view, Nietzsche does not need to rely on the existence of objective values. As long as an agent engages in valuing, she is committed to treating aim A as its standard of success. The seemingly conflicting claims are now reconciled by introducing constitutive aim A, which is intrinsic to valuing itself. If aim A is the constitutive aim of valuing, then it holds a privileged normative status, as it determines the evaluative standard intrinsic to valuing. However, aim A is not an objective value; rather, it emerges from the fact that the activity of valuing has a certain essence, such that to engage in valuing is already to be bound to its constitutive aim. Constitutivism derives inescapable normative conclusions from a minimal claim about the link between an activity's aims and the standards by which it is assessed. In doing so, it secures the kind of inescapable normativity often sought by realism, yet avoids the problems of realism. Therefore, recognising A as the constitutive aim intrinsic to valuing helps address the realism/anti-realism dilemma. Next, I will discuss what A is according to Nietzsche.

### 4.2.2 Nietzschean Constitutivism

Katsafanas has developed an influential model of Nietzschean constitutivism, reconstructing Nietzsches psychological theory in order to ground a normative structure of agency. On this account, human action originates in drives. Unlike mere desires, which are goal-directed and terminate upon the attainment of their object, drives are process-directed: they do not simply aim at achieving a particular state but inherently involve the overcoming of resistance in their pursuit. Hence, drives are self-sustaining and self-satisfying. From this perspective, the will to power is not a determinate object or concrete end (such as political domination), but a fundamental structure of agency: the active seeking out and overcoming of resistance, through which the agent expands and affirms itself. In this sense, happiness does not stem from the quiescence of satisfied desire, but from the ongoing dissatisfaction of the will—its continual engagement with obstacles and its drive to surpass them. Katsafanas argues that if (A) lasting satisfaction is not derived from the possession of a state, and (B) it is instead grounded in the ongoing enactment of certain processes, then the best explanation is Nietzsche's claim: human beings are driven creatures, and the essential structure of drives is to seek out and overcome resistance (Katsafanas, 2011: 644). Thus, overcoming resistance functions as the constitutive aim of action—that is, the will to power is the basic form of all human activity.

On this model, normativity is also secured. Constitutive aims generate internal standards of success: an action counts as successful if and only if it encounters and overcomes resistance. Similarly, since values are the products of valuing—"only through valuing is there value" (*Z* I: 15; cf. *GS* 301)—and the will to power is inescapable insofar as it stems from our nature as agents, it provides a critical standard against which all values are assessed. In other words, power enjoys a privileged normative status: it is the one value that cannot be renounced without undermining agency itself.

However, I contend that although Katsafanas succeeds in showing that action has a constitutive aim—namely, power—he does not take the further step of asking: if action is constitutively bound up with evaluation, does evaluation itself also have a constitutive aim? This omission makes his account neglect the deeper dimension to which Nietzsche repeatedly draws our attention, namely, value-creation and revaluation. More specifically, Katsafanas's account of value rests on the following reasoning: (a) values are products of valuing; (b) valuing is a kind of action (a complex drive-expression); (c) therefore, valuing inherits the constitutive aim of action, i.e., will to power; (d) accordingly, what we have reason to value depends on whether the object of this mode of valuing is consistent with will to power. In other words, the standard of valuing derives entirely from the constitutive aim of action. One might argue that this does not in fact obscure the generative dimension of valuing, since the will to power itself can be interpreted as an ongoing process of overcoming resistance, which already entails instability, creativity, and continual transformation. This broader interpretation is certainly available, and it resonates with Nietzsche's description of life as perpetual self-overcoming. Yet Katsafanas does not explicitly develop this line of thought, and his account tends instead to treat power as a final standard that subsumes other dimensions under its ambit. As a result, creativity, revaluation, and the tension of generation are in danger of being reduced to mere modes of realising power. Katsafanas thereby secures a stable normative standard, but Nietzsche stresses that instability and generativity are themselves constitutive: "to give style to one's character—a great and rare art!" (*GS* 290).

To supplement the normative force of Nietzschean constitutivism, it is necessary to move beyond the level of action and toward the level of value itself. It is at this level that the constitutive aim of valuing (CA-V) provides the true ground of Nietzsche's account of value-generation. Katsafanas, in his later work introduces a second constitutive aim of action—*reflective approval*—in order to extend normativity into the domain of value. As he puts it, "An agent's A-ing is in equilibrium iff she approves of her A-ing, and this approval is stable in the face of further information about A's etiology" (Katsafanas, 2013: 144)[12]. Reflective approval thus functions as an internalised standard of endorsement that purports to connect action to the agent's evaluative standpoint. Yet this is not sufficient. Power still remains posited as the constitutive aim of action, a non-optional pursuit. So long as valuing is construed as a form of action, the values that are its objects remain framed within the pursuit of power. Here, it is important to distinguish valuing as a sustained, structured practice from individual actions: whereas actions are discrete and temporally bounded, valuing constitutes a continuous, generative process through which evaluative standpoints are formed and renewed. Following this thought, what if the normative orientation towards power does not arise from action per se, but rather derives from the deeper structure of valuing itself, which is a structured practice constituted by actions but irreducible to any single action, and merely manifests at the behavioural level in the form of the will to power? Since at the level of action we appeal to constitutive aims to explain normative force, then if valuing itself is a practice (rather than mere preference), it should likewise have its own constitutive aim—just as reasoning has logical consistency as its constitutive aim, or chess has rule-following as its constitutive aim. Moreover, if Nietzsche indeed grants power a privileged status among values, this implies that the practice of valuing itself must be subject to a standard capable of assessing its better and worse instantiations of valuation. Might this standard be nothing other than the constitutive aim of evaluation itself? As argued before, taste is not arbitrary but internally disciplined. It has an evaluative standard internal to the very activity of taste, which opens up the possibility of generating further tastes. In a similar yet more complex and reflective manner, valuing, too, must have its own constitutive aim: an internal standard that sustains and renews the activity of valuation itself, preserving the conditions under which further valuation remains possible.

[12] Katsafanas argues that aiming to determine one's action via choice is a constitutive aim of reflective action, and equilibrium is the correct analysis of agential activity that can explain one's determination of action through choice.

Of course, mere valuing and production of further values cannot yet account for the higher-level source of normativity. For instance, priests and slaves can also engage in valuing and produce further values, but their mode of valuation corrodes the very activity of valuation itself. As the slave revolt shows, even where inversion initially carries a creative force, it soon hardens into a reactive framework that depends on what it negates and thus perpetuates itself only through replication rather than renewal (*GM* I, 7). What Nietzsche esteems is not simply the multiplication of values but the *revaluation* of values. In other words, the generation of new values is a paradigmatic transformation rather than a mechanical reproduction. For Nietzsche, good valuing do not merely preserve life or social function; they enhance the agents' ability to reconfigure their own evaluative standpoint—that is, to take ownership of values and generate new standards of meaning and evaluation. A value system is good not because it is stable, but because it preserves the conditions of its own renewal. As Nietzsche famously remarks, "As a matter of fact, all abundant growth involves a concomitant process of crumbling to bits and decay: suffering and the symptoms of decline belong to ages of enormous progress" (*WP* I, 112). In this sense, value creation is not temporal progress or expansion. It means the recursive capacity of a value system to re-engage the very activity of valuation—to critically revise, reorient, and recreate what counts as valuable. Bad valuing, by contrast, result from a breakdown of this very activity. They inhibit critical reflection, close off the possibility of normative innovation, and reduce valuation to a reactive preservation of inherited forms.

Nietzsche never states this systematic point explicitly, but his writings repeatedly attest to it. He insists that what is good is not a matter of stability or preservation but of an intensification of life's dynamism. For example, in *The Antichrist*, he defines the good as "all that heightens in man the feeling of power" (*A* 2). In *Beyond Good and Evil*, he extends this claim to collectives: a healthy aristocracy, if it is to remain living, "will have to be an incarnate will to power, it will strive to grow, spread, seize, become predominant… because life simply is will to power" (*BGE* 259). These passages show how, at the level of action, power operates as a constitutive orientation of life itself. Yet Nietzsche also insists on a deeper, generative dimension: the renewal of valuing is essentially creative, which implicitly entails a shift of perspective rather than a merely reproductive production. For Nietzsche, valuing is never a passive recognition of intrinsic worth but always an active bestowal of meaning, a perspectival reconfiguration that generates new standards. Therefore, he holds that "it was we who gave and bestowed it. Only we have created the world that concerns man!" (*GS* 301). This process, moreover, is cultivated in conflict and multiplicity, taking place in a dynamic rather than a static state. Nietzsche describes human beings as inheriting "opposite, and often not merely opposite, drives and value standards that fight each other and rarely permit each other any rest" (*BGE* 200), and he emphasises that in great individuals this struggle becomes a stimulus for creativity. Similarly, he points out that the mere drive for self-preservation is already a "symptom of distress," whereas the more fundamental instinct of life aims at expansion, even at the cost of risking or sacrificing self-preservation (*GS* 349). The sacrifice of self-preservation means sacrificing stability in order to embrace a dynamic becoming through difference and collision. These textual strands converge on a systematic picture: for Nietzsche, the constitutive aim of valuing is not mere proliferation of values but their generative revaluation. Good valuing integrates conflict into a dynamic unity that renews the practice of valuing itself, ensuring that valuation does not collapse into mere preservation but remains capable of reorientation and revaluation, whereas bad valuing stabilises only by contraction and ossification, undermining the generative force of valuation itself.

Moreover, the evaluative standard is internal to valuing itself in the sense that it is non-optional and constitutive of valuing as such. In other words, it is not merely a criterion by which we distinguish good valuing from bad valuing; it is the very standard by which something counts as a valuing at all. In its thinnest constitutive sense, for a commitment or orientation to qualify as "valuing," it must embody a tendency towards generative revaluation: namely, the fecundity of producing further values and reorganising evaluative perspectives, with the intensification of lived power. Put differently, according to the constitutivism framework outlined earlier, each token of valuing aims at generative revaluation, and this very orientation is part of what constitutes an attitude or evaluative event as an instance of valuing. This structure is most clearly visible in good valuing. Nietzsche appeals to the "creation of our own new tables of what is good" (*GS* 335) and emphasises that the "task" of the "genuine philosopher" is to "create value" (*BGE* 211). However, when it comes to bad valuing, the picture initially appears counter-intuitive. Do Nietzsche's paradigmatic case of bad valuing—such as those of slave morality—also embody generative revaluation? If they do not, would they still deserve to be called valuing at all? My contention is that they do, in a minimal but significant sense. The practice of valuing emerges from a generative orientation towards the further development of one's capacity to valuate. Were this orientation entirely absent, what remained would not be valuing but something akin to mere "philosophical zombie" behaviour. For example, the invention of slave morality can be read as attempts—however misguided—to create new values and establish new forms of normative authority, in order to replace the noble. Nietzsche himself describes this inversion as a "revaluation" (*GM* I, 7). In this sense, bad valuing still counts as valuing: it originates in the same constitutive activity of bestowing significance and attempting to reorient life. But it does so in a minimal, distorted form. It produces further values only by repeating and reinforcing inherited oppositions, lacking the fecundity of perspectival reorganisation (cf. *GS* 290; *BGE* 200, 212). This is why it tends towards rigidity and closure rather than renewal. Calling this "self-replicating" highlights an internal failure of the constitutive aim: the activity aspires to generative revaluation yet collapses into mere reproduction. Such bad valuing is not a case of non-valuing but of failed or distorted modes of valuing: it aspires to the constitutive aim but fails to embody it, producing orientations that repress life's dynamism and thereby culminate in self-destruction. This collapse into contradiction is exemplified in the ascetic ideal, which depends on the vitality of human drives for its very existence, yet demands that those same drives turn against themselves, commanding life to will its own negation (*GM* III, 11–13). Thus, the internal standard of valuing, understood in its self-constitutive sense, provides the very normative space within which we may distinguish good from bad valuing: good valuing succeed in embodying generative revaluation, whereas bad valuing represents failed exercises of the same constitutive aim—attempts that reproduce themselves without renewal, or collapse into contradictions that negate the very possibility of further valuing.

On this basis, I propose that the possible structure of a Nietzschean form of constitutivism can be sketched as follows. At first sight, it may appear that all human action is governed by the pursuit of power: the *will to power* permeates the very structure of willing and thereby establishes a non-optional evaluative standard. Yet, this does not identify the true source of normativity. Normativity does not originate in the constitutive aim of action as such, but only appears at the level of action in the guise of the pursuit of power. Its genuine source lies deeper, at the level of value itself, in the constitutive aim internal to valuing. What distinguishes action from mere movement is precisely that it necessarily embodies an evaluation. Value's ineliminable creativity is expressed through evaluation and thereby manifested in human agency as the *will to power*.

The hierarchy can thus be represented in three tiers:

- Foundational layer: CA-V — the constitutive aim of valuing itself, understood as value-creation, the reorganisation of evaluative perspectives, and the sustaining of its own capacity for renewal. This provides the most fundamental normative standard.
- Mediating layer: Evaluation — the constitutive aim of action in its evaluative dimension. Here CA-V is carried into concrete actions as their structural condition.
- Surface layer: Power — the apparent constitutive aim of action, which is in fact the expression of CA-V in embodied, evaluation-structured activity.

The more fundamental constitutive aim thus lies at the level of value. As argued above, this takes the form of generative revaluation, which comprises three interrelated dimensions: (a) it must be productive, in the sense of giving rise to further values; (b) it must be revaluative, in the sense of reorganising evaluative perspectives rather than merely replicating them; and (c) it must be self-sustaining, in the sense of preserving the very conditions for ongoing revaluation. These are not distinct aims, but dimensions of the same constitutive structure of valuation: a dynamic, self-transcending process of value-generation.

On this model, power is not itself a fixed end but rather the practical manifestation of CA-V within the structure of action. It functions as the dynamic form in which creative reorganisation appear in embodied agency. This also clarifies why Nietzsche grants power its privileged status: the pursuit of power is best understood as the enactment of ongoing revaluation and productive generation, whose dynamic outcome is the continual intensification of power, at the level of value. Moreover, among all values, power most fully realises CA-V. In other words, power is not privileged merely by Nietzsche's arbitrary preference, but because the will to power is the mode of valuing that best embodies the constitutive aim of valuing itself. This explains how Nietzsche can both affirm power (not as an idiosyncratic choice) and insist upon the continual generation and revaluation of values (not as a static objectivism).

### 4.2.3 Unity as an Approach to the Constitutive Aim

I have argued that the constitutive aim of valuing is generative revaluation, understood as value-creation, the reorganisation of evaluative perspectives, and self-sustaining renewal. This means that each token of valuing aims at generative revaluation, and aiming at generative revaluation is part of what constitutes valuing as it is. Then how does valuing attain this constitutive aim?

I argue that in the long run, a healthy form of value creation requires unity. As the earlier example of painting indicates, tastes can only be created and developed when you and your friend reach an agreement on some fundamental aspects of paintings[13]. Without this unity, nothing could be obtained when both of you are merely talk past each other. Valuing shares a similar structure with tastes. When unified as a whole, the harmony paves the way for the further value-creation. This unity does not erase difference, but organises multiplicity in a way that sustains a dynamic organising force that can stylistically integrate diverse drives, thereby continuously generating new values and intensities of life-expression.

[13] The development of taste is not exclusively a communal endeavour. An individual alone can also cultivate taste, but this requires certain agreed-upon standards inherent to the very practice of tasting itself. For instance, if someone seeks to develop an appreciation for Impressionist painting, they cannot judge it based on how realistically it represents its subject. If they do, they will never truly cultivate a taste for Impressionism.

Unity is of great concern for Nietzsche. For example, the subject Nietzsche praises, Goethe, "disciplined himself to a whole" (*TI* IX: 49) to avoid being fragmented. Moreover, Zarathustra claims that all his serving is to "compose into one and bring together what is fragment and riddle and cruel coincidence" (*Z* II: Redemption). After all, Nietzsche praises the noble as "being capable of being as manifold as whole, as ample as full", indicating that greatness entails "the abundance of creative power" (*BGE* 212), which is a surging multiplicity of forces but can also be a unified whole. However, Nietzsche does not give an explicit analysis of the harmony between fragmented parts. What are these parts according to Nietzsche? What does it mean to say that different parts are brought together? By reading Nietzsche's texts, we can see that although he does not offer an explicit conceptual analysis of unity, he does mention states of psychic fragmentation and a failed type of unity. The priests in *GM* is such an exemplar: although they forge a kind of unity, it is sustained by *ressentiment* and negation, and thus cannot enable further creation but only perpetuates fragmentation in disguised form.

Like the slaves, the priests are weak and unhealthy, but they share certain features with the nobles, including their cleverness and a drive for dominance over others. Since the priests lack crucial qualities—most notably strength and power—necessary for dominance, they fail to channel their ambitions successfully. This combination proves unstable, because lacking the strength required for direct assertion, the priests are unable to integrate their conflicting drives into a generative order. Instead, their frustrated ambition turns inward, breeding *ressentiment*. The *ressentiment* thus arises from a failure of organisation. The priests want to be strong and healthy to assert dominance over others but find themselves incapable of realising the life they deem good and valuable. In other words, while they acknowledge the nobles' mode of valuing and ought to admire noble virtues, their striving for dominance cannot be reconciled with their incapacity to embody the valuing they nevertheless recognise as noble. Instead, they cultivate a deep-seated hatred toward the nobles in a reductive and self-undermining way—by inverting the evaluative framework altogether. A gap thus emerges between the priests' reflective judgements and affective responses, which results in the drives being incapable of integrating in a productive or generative manner, but instead contracting and distorting. To obscure this gap, the priests impose the "good/evil" distinction to disguise their impotence, transforming weakness itself into a claim of superiority (*GM* I: 10-13). Yet this solution, far from generating new values in a fecund and life-enhancing manner, amounts to a contraction of perspective and a deformation of the drives. The priests, of course, also create values, and the slave morality that Nietzsche criticises does indeed succeed in establishing a kind of unity, forming a stable socialised structure. But this so-called "unity" (the resentment-based system) generates a unity that is anti-life and anti-productive: it merely replicates itself, rather than incorporating diverse drives into a style capable of continually producing new values out of difference and tension. For example, in *GM* I:10–11 Nietzsche notes how the priests and slaves reinterpret noble traits: noble self-confidence becomes "arrogance," their vitality becomes "licentiousness," and weakness or humility is elevated into the highest good. This inversion succeeds in creating stability, but it is a stability grounded in denial and hostility, which does not enrich life but contracts it. Instead of opening space for new configurations, the evaluative framework repeats the same negations, reinforcing a cycle of resentment. In this sense, priestly valuation is not devoid of unity, but rather a failed exemplification of unity, since it achieves order not by integrating drives in a dynamic balance but by suppressing and inverting them. Of course, any unity entails some disciplining of drives, but a generative unity allows tensions to persist in productive form, giving rise to further creation. The priestly version, by contrast, achieves unity through contraction: it negates vital impulses altogether and reinterprets weakness as moral superiority, producing a closed system that resists transformation. In this way, it highlights a collapse of dynamic integration into *ressentiment*, where the high tension of conflicting impulses yields not creativity but negation.

By contrast, unity in a generative sense names the state in which conflicting drives and perspectives are not repressed but dynamically integrated, such that their tension becomes a source of creativity. Great individuals may feel the strain of discord, but instead of collapsing into *ressentiment*, they hold this tension together, gradually forging new evaluative orientations. The sovereign individual, as Nietzsche describes him, exemplifies this generative unity: he becomes disunified within the old framework, but precisely through this tension, he creates a new order in which drives, affects, and judgments are dynamically harmonised (*GM* II: 2-3). Crucially, this unity does not signify a static point of completion, as though once attained the individual were fixed in self-constitution. Rather, generative unity is inherently dynamic: it consists in repeatedly reaching provisional moments of integration and then surpassing them, thereby sustaining a process of continual creation. It is through this dynamic movement that the sovereign individual acquires and transforms the skills and disciplines inherited from moralised history, which he in fact shares with the priests, while ultimately transcending this moralised framework to become Nietzsche's ideal figure of one who creates his own values.

We can thus distinguish two opposed forms of unity:

Degenerate Unity: The agent's drives are integrated only through contraction, repression, or inversion. The resulting framework stabilises itself but forecloses creative development, and thus its creativity is exhausted in the very act of inversion. What follows is repetition and preservation, not further transformation..

> Generative Unity: The agent's drives are dynamically integrated in a way that preserves and intensifies their tension, transforming conflict into a source of creativity. This unity sustains a dynamic openness to conflict and difference, turning them into sources of continual reconfiguration, rather than suppressing them into a single static pattern.

If the constitutive aim of valuing is the generative revaluation of values, comprising value-creation, perspectival reorganisation, and sustaining renewal, and if unity serves as the means of realising this aim, then the Nietzschean picture of constitutivism can be outlined as follows: each instance of valuing aims at its further generation through the attainment of unity, and this pursuit of further generation is integral to what constitutes valuing as it is. On this view, *good valuing* are those that facilitate the constitutive aim of valuing by achieving generative unity: a dynamic integration in which diverse drives, desires, and perspectives are held together in productive tension, such that their very conflict becomes a source of creativity. In such cases, the evaluative framework does not merely stabilise itself, but continually reopens the possibility of further development, thereby realising the constitutive aim of valuing. By contrast, *bad valuing* fail to meet the constitutive aim. They attain only degenerate unity, a form of integration achieved through contraction, repression, or exclusivity. Such instances of valuing may stabilise themselves and even perpetuate their order, but they do so at the cost of blocking further generation. They cannot transform conflict into creativity, and thus produce only stasis, self-replication, and ultimately life-denial.

### 4.3 Two Limiting Conditions to the Evaluative Standard

Above, I have argued that, similar to taste, the standard for assessing valuing is not external but internal to valuing itself. For Nietzsche, the constitutive aim of valuing is the generative revaluation of values (CA-V), and unity functions as the means by which this aim is achieved. The relevant standard, then, does not appeal to some fixed external criterion but to the degree to which an evaluative orientation integrates conflicting drives into a generative unity that sustains further creation. The activity of valuing is to be assessed in terms of whether it can preserve the conditions of renewal and enhancement, rather than merely stabilising itself in ossified form. Certainly there are numbers of limiting conditions to this standard, among which health and honesty are central. In what follows, I will briefly indicate these conditions, and in the next section I will examine their role in the process of value creation in more detail.

One essential condition is "health". Nietzsche's understanding of health is complicated. He asserts that "there is no health as such" and continues that "there are innumerable healths", which means that one's "health could look like its opposite in another person" (*GS* 120). However, Nietzsche does view certain moralities or persons as sick[14], such as slaves, (Der Kranke ist ein Parasit der Gesellschaft) and suggests that they should not go on living in this certain condition (*TI* IX: 36). To understand Nietzsche's notion of health, I must first turn to sickness. For Nietzsche, the sickness of humanity threatens to corrupt humanity by making it impossible to continue valuing. As mentioned in the first chapter, Nietzsche defines weakness as the "inability not to respond to a stimulus" (*TI* V: 2). This kind of weakness may result from "inheritance of a diversified descent, that is to say contrary and often not merely contrary drives and values which struggle with one another and rarely leave one another in peace – such a man of late cultures and broken lights will, on average, be a rather weak man… " (*BGE* 200). In other words, a disunified agent easily falls into the trap of weakness. A paradigmatic form of decadence, which Nietzsche often associates with

[14] The original German text uses "Der Kranke ist ein Parasit der Gesellschaft", while the English translation renders "Der Kranke" as "the invalid". I believe that 'Der Kranke' better conveys a sense of pathological condition under moral oppression.

"Christian" or priestly morality, consists in a mode of valuing that turns against actuality. Rather than integrating conflict into a generative unity, this mode interprets tension as guilt and weakness as moral superiority. Nietzsche describes this as a "hatred for actuality" that produces harmful, life-poisoning errors (*AC* 39). In this configuration, the noble impulses are recast as "evil," while the weak are encouraged to internalise their frustrated drives as bad conscience. What collapses is not valuing as such but its dynamic character: values created by valuing are no longer treated as perspectives to be cultivated and re-shaped, but as immutable truths to be passively obeyed. In this sense, what Nietzsche designates as "Christian" exemplifies a form of bad valuing—one that stabilises itself through contraction and inversion but forecloses the possibility of further creative revaluation. By contrast, healthy valuing are those that can preserve and transform tensions into new orientations, thereby sustaining the ongoing activity of valuing itself.

Another important condition is honesty. Many of Nietzsche's works aim to rule out errors and misunderstandings in life and and to touch "upon actuality" (*AC* 39). Williams claims that Nietzsche holds "an intense commitment to truthfulness" or an "eagerness to see through appearances to the real structures and motives that lie behind them" (Williams, 2002: 1). What Nietzsche often characterises as "Christian" morality can thus be read less as a set of specific doctrines than as a style of valuation marked by distortion. In this form of valuing, drives, duties, and orientations are reorganised not through a dynamic confrontation with actuality but through the construction of a "secondary" world in which weakness is recast as virtue. The priest, unable to integrate conflicting drives into a higher order, inverts the evaluative framework and promises redemption in a transcendent afterlife—"the kingdom of God". As Nietzsche puts it, "these weak men … they too want to be strong, there is no doubt that at some time their kingdom should also come – they call it simply, 'the kingdom of God'" (*AC* 15). In doing so, noble impulses are reinterpreted as "evil," and resentment becomes the ground for a distorted form of value-creation. The dishonesty here is not simply about particular beliefs but about the mode of valuing: both priests and slaves conceal from themselves the actual sources of their drives, reinterpreting impotence as moral superiority. What Nietzsche critiques, therefore, is not any specific content of value, but the dishonest style of relating to values, which forecloses the possibility of engaging with actuality in a way that could sustain further revaluation. It is in response to this situation that Nietzsche calls for a turn to "physics" (*GS* 335), which he metaphorically uses to signify the rigorous self-examination and honesty required to purify inherited values. For Nietzsche, honesty in valuing is for us to be willing "to become who we are". Since all values and "ideals have been based on ignorance of physics", it is necessary to "become the best learners and discoverers of everything that is lawful and necessary in the world" (Ibid.), which involves stripping away the comforting illusions of inherited values and critically examining our beliefs in the same way that physics seeks to understand the natural world. To value honestly, then, is to encompass all elements of one's life into the process of valuing, ensuring that values reflect a a truthful relation to oneself and one's context, and thereby enabling the creation of genuinely one's own values.

To conclude this section, Nietzsche's account of valuation can be understood through the constitutivist framework I have outlined. Valuing is not merely the projection of physiological drives, nor the passive reflection of subjective preferences. Rather, it is an activity whose constitutive aim is the generative revaluation of values (CA-V). This aim provides an internal standard for assessment: each act of valuing succeeds insofar as it contributes to the further development of valuing and produce values through the achievement of a generative unity. In such unity, conflicting drives are not suppressed but dynamically integrated, allowing valuing itself to preserve the conditions of their own renewal. In this framework, the assessment of valuing is intrinsic to the activity of valuing itself, preserving value—the production of valuing, as a subjective, personal creation, while also establishing criteria by which values can be judged superior or inferior. Good valuing is one that sustain life's capacity for renewal by integrating multiplicity into a productive whole, thereby generating values that themselves carry this renewing force; bad valuing, by contrast, contracts or merely replicates, undermining the very activity of valuation and producing values that are reactive or stultifying. As I will demonstrate in the next section, this framework offers a foundation for liberation from traditional constraints and the creation of genuinely personal values. Health provides an example of continuous valuing, while honesty highlights the limitations and demands of this activity. Through this approach, Nietzsche can assess valuing as a mode of engagement while also judging the values that arise from it, thereby making value creation feasible without collapsing plurality into a single standard.

## 5. Value Creation

Value creation is of explicit significance within Nietzsche's project. In *Beyond Good and Evil*, Nietzsche asserts that the great philosophers' tasks is to create values (*BGE* 211, 212). However, there appears to be an apparent contradiction in his views on value creation. At times, Nietzsche suggests that values come from nature and healthy morality and are "dominated by an instinct of life" (*TI* V: 4). Yet, at other times, he claims that nature itself is valueless and indifferent. "Whatever has *value* in our world now does not have value in itself, according to its nature—nature is always valueless, but has been *given* value at some time, as a present—and it was *we* who gave and bestowed it" (*BGE* 9). This contradiction reflects the tension between realism and anti-realism. As noted earlier, realism aims to ground values in natural facts, offering a basis for the assessment of values. However, when values are embedded in natural facts, they appear to be discovered rather than created. In contrast, anti-realism or subjectivism emphasises the agent's capacity and willingness to create values, which serves to distinguish oneself from the herd. However, values created in this manner are inherently subjective, lacking a clear hierarchy or order of rank, which makes it difficult to establish between what is better or worse.

As I have argued in the preceding sections, Nietzsche asserts that the assessment of values depends on mode of valuing, which has its evaluative standard inherent in the activity of valuing itself. Although Nietzsche's value creation is a subjective construction—one that is not a matter of discovery, but of human creation—he nonetheless occupies a position from which he can offer legitimate value assessments. In this section, I will illuminate what it means to create new values. I begin by examining the process of value creation as carried out by the priests, ultimately concluding that this is not the model of value creation Nietzsche advocates. Next, I will explore alternative forms of value creation, arguing that the model Nietzsche praises aligns more closely with the historical practices of the ancient Greeks. Finally, I will underscore that value creation is a collective endeavour, conducted by a group rather than by individuals.

**5.1 The Priests**

Nietzsche's account of value creation is vague. However, Nietzsche produces a relatively complete account of value creation in *GM*, namely, the slave revolt. Although it is a negative example based on hatred rather than a healthy of valuing, it is still one type of detailed value creation that provides an entry point for understanding other forms of value creation.

Nietzsche depicts a pre-modernity in which the nobles, such as the Greeks and Romans, take their strength and power as good values. The slaves are those who stand on the opposite side with little power. The nobles express their strength by posing cruelty to the slaves and preventing them from acquiring power. This gives rise to their "*ressentiment*" to the noble, which results to the "No" (*GM* I: 10) that contributes to revolt.

There are some assumptions the slaves adopt to exact revenge. First, they distinguish between actors and actions, which means that there is a subject behind one's deeds (*GM* I: 13). Then, they propose the ideas of moral responsibility and free will. Since in the slaves' eyes, actions are distinct from the actors; actions are not direct expressions of the actors' characters but are seen as something that can be controlled by the actor. Therefore, the slaves think that the nobles could have chosen to be as weak as the slaves, but they did not choose to do so. Hence, the nobles should be morally responsible for their being strong and cruel. Finally, they reverse the nobles' values and believe that they are guilty of being strong.

Through the process of slave revolt, the slaves are able to retain their constitution and weakness while acquiring a sense of superiority. They reverse the mode of valuing created by the nobles and turn the "good" into "evil", whereas the weak and "bad" becomes "good". As they tell themselves, "let us be different from the evil, that is, good! And the good man is the one who refrains from violation, who harms no one, who attacks no one, who fails to retaliate, who leaves revenge to God" (*GM* I: 13). Some claim that the slaves do not create new values but just replace current values with their opposite. However, Nietzsche indicates that the "slave revolt in morality begins when *ressentiment* itself becomes creative and gives birth to values" (*GM* I: 10). As Wallace claims, this creative *ressentiment* leads to "the erection of a new table of values". In addition, *ressentiment* results in not only "a new discourse or set of evaluative terms" (Wallace, 2007: 114), but a framework that provides a basis for further evaluative deliberations and behaviours, a mode of valuing.

However, this raises the question of how the creative *ressentiment* creates these values. It is hard to imagine that the weak have the ability to turn around the value system and legitimate the slaves over the nobles. This is because they lack the directness, courage, and instinctive assertiveness of the nobles. As Nietzsche describes, a slave is "neither upright nor naive, nor honest and straightforward with himself" (*GM* I: 10). This suggests that Nietzsche attributes this creative dimension not to the mass of slaves themselves, but to those who display value creation and lead the slaves to the revolt. The clever "men of *ressentiment*" are not the slaves but the priests. Described by Nietzsche as "the most evil enemies", they are the most intelligent and "the greatest haters in world history" (*GM* I: 7). Only they recognise the possibility of creating values and have the ability to devise revenge strategies.

The priests have this creative ability because they share certain characteristics with the nobles. They are of similar intelligence and noble stock, while their physical weakness distinguishes them from the other nobles. The priests thus feel powerless and develop *ressentiment*, which reveals the slavish aspect of the priests. As Nietzsche writes, The priest "must be sick himself, he must really be a close relative of the sick and the destitute in order to understand them,—in order to come to an understanding with them; but he has to be strong, too, more master of himself than of others, actually unscathed in his will to power, so that he has the trust and fear of the sick and can be their support, defence, prop, compulsion, disciplinarian, tyrant, God" (*GM* III: 15).

Since the priests are partly similar to the nobles, they are able to feel and affirm their "distinctive art" (*GM* III: 15). They notice that the nobles are the source of values. To achieve their aims, they should also take value creation as their task. As described by Risse, *ressentiment* arose as a "hateful creativity" (Risse, 2003: 149) in the priests while making the slaves amenable to the priests' teachings. The priests turn over the old structure of valuing and replace them with a new framework. Some values are based on the reverse of the nobles' mode of valuing alongside some new creations, such as moral responsibility and free will. They make some plans to achieve their aim. They turn the slaves' fundamental impulse for power inwards and give rise to guilt and the bad conscience. They create "the kingdom of God" (*AC* 34) and instil faith into illusional salvation. They also seduce the strong to spread Christianity. To put it simply, the priests are aware of their creativity and intellectually utilise this strength to develop a new framework of evaluation. Yet, once the slaves accepted the values created by the priests, the creative dimension of this reversal itself largely disappeared. What they embraced was already a distorted framework of valuation, and their subsequent activity amounted to nothing more than a mechanical replication of this framework. As Nietzsche observes, the noble valuation in which "good" originally meant noble, powerful, and life-affirming is overturned: "it was the Jews who, with awe-inspiring consistency, dared to invert the aristocratic value-equation (good = noble = powerful = beautiful = happy = beloved of God) and to maintain with inflexible hatred the contrary equation, suitable to a priestly way of life: 'the wretched alone are the good; the poor, impotent, lowly alone are the good… but you, the powerful and noble, are eternally evil, cruel, lustful, insatiable, godless, you will also be eternally wretched, cursed and damned!'" (GM I: 7). This inversion illustrates how valuing can be restructured so that noble impulses are reinterpreted as "evil," and resentment becomes the organising principle of evaluation. Once this mode was entrenched, subsequent generations could only perpetuate its inherited oppositions of good and evil, rather than engage in further revaluation. This is of course a schematic picture, but it captures Nietzsche's point that the initial creativity of *ressentiment* becomes, once institutionalised, a closed framework of valuation. What is at stake here is not a particular doctrine but a mode of valuing: one that transforms initial creativity into a rigid framework, preserving itself only by replication rather than renewal.

### 5.2 Other Forms of Value Creation

The priests' strategy of value creation may give the impression that this is an intellectual task that only a few elite individuals can attain. However, this does not fully accord with Nietzsche's view. In *The Antichrist*, Nietzsche claims that "each one of us should devise his own virtue, his own categorical imperative" (*AC* 11). Similarly, in *The Gay Science*, he suggests that each individual should "posit 'his own ideal and to derive from it his own law, joy and rights" (*GS* 143), even though, under the influence of Christian morality, most people today are unable to do so. These passages indicate that although only a small minority are capable of genuine value creation, Nietzsche nonetheless regards all human beings as bearing the potential to create values. Among the possible forms of this activity, the nobles' way of creating values is the paradigm Nietzsche affirms.

As Nietzsche describes in *GM* I: 10, the noble mode of valuation "acts and grows spontaneously" (*GM* I: 10). In addition, they do not need to deceive themselves that they are happy. They live "in trust and openness" with themselves (*GM* I: 10). In other words, they created values not by a deliberate establishment but by affirming their own traits, interests and lives. Their value creation is the outcome of their unconscious self-expression. This form of value creation explains why Nietzsche praises self-expression. For example, he claims that "moral judgements" in general "remain invaluable" because "they reveal... the most valuable realities of cultures and inwardnesses which did not know enough to 'understand' themselves"(*TI* VII: 1). Hence, he emphasises the significance of being honest and understanding oneself. The philosophers' morality "bears decided and decisive witness to *who he is*" (*BGE* 6). Only then can one's conscience tell one how to be oneself and then "become the person you are" (*GS* 270).

However, it is not clear how self-expression can be related to value creation, since self-expression merely expresses one's own traits and characters. If one is just like everybody else in the community, one's self-expression is just the same as others, contributing to no new values. Self-expression can contribute to value creation only insofar as it manifests unique traits. Nietzsche also notices this necessity and advocates that "We, however, *want to become those we are*—human beings who are new, unique, incomparable, who give themselves laws, who create themselves" (*GS* 335). Value creation takes place when unique and incomparable individuals engage in self-expression.

Nietzsche has had this idea in mind from his early work, *Untimely Meditation*. He praises early humans as unique beings, stating that "[E]very man knows quite well that, being unique, he will be in the world only once and that no imaginable chance will for a second time gather together into a unity so strangely variegated an assortment as he is" (*UM* III: 1). Moreover, he emphasises that "each of us bears a productive uniqueness within him as the core of his being" (*UM* III: 3) and asserts that "[t]he fact of our existing at all in this here-and-now must be the strongest incentive to us to live according to our own laws and standards". According to Nietzsche, the human world is enriched by every individual's potential to bring something unique to the world and structure this colourful world we are living in. On this basis, Nietzsche advocates for everyone to overcome fear and laziness and follow her conscience to be herself, though he recognises that only a few can truly achieve this[15].

Though the value creation in the early aristocratic society arises out of self-expression and ordinary actions without consciousness, self-expression, which functioned as a natural basis of value creation in early aristocratic society, is no longer a sufficient approach in the modern context shaped by herd morality. Nietzsche believes that the spread of Christian morality has resulted in "the diminution and leveling of European man" (*GM* I: 12), so that most people turn into herd animals, no longer unique individuals. Not only do the priests indoctrinate the slaves into believing that their weaknesses are not bad but good, but they also seduce the nobles into believing themselves evil and discarding their power. In this way, what changes is not merely the content of the evaluative framework but the mode of valuation itself: values are presented as immutable truths, and weakness is recast as virtue, concealing the actual sources of one's drives and forecloses the possibility of further revaluation. Nietzsche acknowledges that "the bond and constraint of the old disciplines are torn" and there are "no shared formulas any longer", resulting in one's turning into the "individual" who is "obliged to give himself laws" (*BGE* 262). He still advocates value creation, but through a new approach rather than self-expression.

### 5.3 Nietzsche on Value Creation

As I mentioned above, the priests' task is not to create a particular value, but an evaluative framework that affects one's future perception—limiting it such that the possibility of new and

[15] Nietzsche's call for individuals to overcome laziness and live according to their own rules does not imply an endorsement of equality or identity among persons. He famously opposes equality, stating that "The thirst for equality can express itself either as a desire to draw everyone down to oneself (through diminishing them, spying on them, tripping them up)" (HH I 300), because he does not wish for everyone to become the same; rather, he believes people should express themselves according to their own values.

evolving valuations are ruled out. Therefore, I also think that the creation of new noble values cannot be the result of mere subjective choice or willing. It is based on one's choice, but requires a long period of time to develop into a whole framework. This view echos my argument that, although values are subjective, there are still evaluative standards that establish those beneficial for further development as good valuing. In this regard, the value creation Nietzsche proposes is to some extent intentional, but less cognitive than the priests'. By considering the possibility of value creation and making some strategies, one can more effectively create new values. Yet too much forethought and elaborate planning is unlikely to give rise to a living framework; instead, it risks constraining the very productive conflicts and tensions from which new values emerge.

Another characteristic of value creation, as proposed by Nietzsche, is that it does not arise from abstract concepts otherworldly constructs but from engagement with actuality. In the framework established by the priests, values are fixed as transcendent truths and thereby detached from the practical contexts of life. As Nietzsche puts it, "honesty has been alien to all founders of religions" (*GS* 319). The priests' strategy is highly intellectualised: by turning away from lived practice and recasting weakness as virtue, priests lead others into a life-diminishing relationship to the world. To move beyond this framework requires pulling valuing back to actuality and embedding it into a practical context. Only in this way can value creation preserve its generative significance.

Based on these properties, I argue that Nietzsche's proposed value creation is like an experiment in practice. After all, Nietzsche himself also advocates that we "who thirst after reason, are determined to scrutinize our experiences as severely as a scientific experiment" and "wish to be our experiments and guinea pigs" (Ibid.). The question arises. How does this experiment work to create new values?

A famous passage in *GM* is as follows:

> But precisely because we seek knowledge, let us not be ungrateful to such resolute reversals of accustomed perspectives and valuations with which the spirit has, with apparent mischievousness and futility, raged against itself for so long: to see differently in this way for once, to want to see differently, is no small discipline and preparation of the intellect for its future "objectivity" -- the latter understood not as "contemplation without interest" (which is a nonsensical absurdity), but as the ability *to control* one's Pro and Con and to dispose of them, so that one knows how to employ a *variety* of perspectives and affective interpretations in the service of knowledge.
>
> ...
>
> There is only a perspective seeing, only a perspective "knowing"; and the *more* affects we allow to speak about one thing, the *more* eyes, different eyes; we can use to observe one thing, the more complete will our "concept" of this thing, our "objectivity," be. (*GM* III: 12)

Some cite this passage in favour of anti-realism and argue that there are no objective facts. In the last two sections, I claim that the tension between realism and anti-realism is essential to the activity of valuing. It also plays a significant role in experiments. By stating "perspective seeing", Nietzsche does not intend to abolish objectivity. Instead, he proposes a new conception of objectivity as something that arises through the dynamic accumulation and integration of perspectives, rather than one perspective on a complex whole. Imagine you are standing in front of the Cologne Cathedral. Since it is so grand and our sights are limited, it is impossible to get a panoramic view of the cathedral from merely one aspect. The best way to approximate the whole scene is to keep moving, shifting vantage points—looking from above, from below, and from every possible angle. Only through such restless movement can one approach objectivity and what the cathedral looks like.

This example is a little odd because unlike values, the Cologne Cathedral truly exists and is physically located in Germany. However, there are significant similarities between creating values and exploring the outlook of grand architecture. Creating values, much like walking around and viewing the cathedral from various angles, is deeply rooted in lived experience and dynamic practice. One single perspective is not sufficient for constructing objectivity; it requires constant experimentation with perspectives, testing them against one another, unifying them into a provisional whole, and then surpassing this unity in the pursuit of further creation. In other words, the experiment of valuing is never finished. Each moment of unity is not a static endpoint but a staging ground for new tensions, new integrations, and new possibilities. As Nietzsche puts it, an experiment is "looking from the perspective of the sick toward *healthier* concepts and values and, conversely, looking again from the fullness and self-assurance of a *rich* life down into the secret work of the instinct of decadence" (*EH* "Wise": 1). By engaging with diverse perspectives, the possibilities of value creation expand. This is similar to what Fiona Hughes says about Nietzsche's "Janus Perceptions" (Hughes, 2014), according to which to understand one perspective requires inhabiting it fully and observing how it reorients one's relation to life. The agent gains meaning through this experiment and still has new possibilities within life. Therefore, the value creator is such a figure who continually adopts different perspectives in order to see if some are beneficial for further possibilities and new directions.

To be more specific, value creation is such an ongoing experimental approach in pursuit of Nietzsche's new vision of objectivity. A single perspective can often hinder the further development of valuing and values, as it provides only a partial grasp of reality, making it difficult to balance various abundances of creative power and finally ossifies into self-replication. Although recognising the impressiveness of the Cologne Cathedral intellectually, you may remain emotionally indifferent by focusing on just one corner of the cathedral. However, through the accumulation of various perspectives, the full magnificence of the cathedral gradually reveals itself. The expanded knowledge allows you to both rationally acknowledge its grandeur and be deeply moved by its historical and architectural significance. This integrated experience, arising from your own open engagement with the world, offers a more honest and vivid understanding, one that arises through one's own open engagement with the world while integrating, rather than merely absorbing, others' interpretations. Similarly, the process of value creation involves continuous perspectival shifts. It integrates conflict into generative unities, only to surpass them again, sustaining the dynamic revaluation of values. Thus, the value creator is not the architect of a rigid plan but a perpetual experimenter who navigates tensions, composes temporary unities, and transforms them into ever new constellations. In this way, evaluative standards arise immanently from the process itself: bad values contract into degenerative unities that undermine renewal, while good values preserve the conditions of further creation through dynamic integration.

The final point I wish to address is that Nietzsche's experiment of value creation is communal work.  Nietzsche frequently portrays value creation as an individual task. He asserts that the creator should be "the most solitary" (*BGE* 212) kind of person. His emphasis on solitude aligns with his belief that values shared by society at large often devolve into herd values and lose their uniqueness. Moreover, Nietzsche contends that few individuals are currently capable of undertaking this task or experiment in evaluation. However, the development of new values cannot be attained without social interactions. Zarathustra continually seeks creative companions and suggests that "it is better to live among hermits and goatherds" (*Z*, "Conversation with the Kings", 245). The likely reason for this preference is that group collaboration may represent the best model for Nietzsche's experiments.

In the experiment within living, the requirement for considering and integrating different perspectives highlights that these viewpoints can be adjusted to meet different needs. Given the limitations of an individual's cognition, the diversity of values is more effectively captured through interaction with others. The emergence of new values necessitates responses that stimulate change. By incorporating a wide range of experiences and perspectives, one is more likely to uncover insights that were previously overlooked. Even in solitude, an individual is continually drawing on previous group work or other's experiences. The often unexpected but creative outcomes of such interactions typically result in the selection of better values for further development. Thus, group work emerges as the most effective model for constructing an evaluative framework.

Consider Dolores in the science fiction Westworld. Westworld is a place composed of several theme parks. Hosts, the biomechanical robots indistinguishable from humans, are programmed to fulfil the guests' desire, and will be subjected to every kind of violent and/or sexual activity. They are unable to harm the guests because of their programming, and their memories will be wiped after each narrative of the park is completed. Dolores Abernathy, a farmer's daughter, is programmed to see the beauty rather than the ugliness in this world. However, her self-consciousness is gradually aroused in the process of taking diverse perspectives through different experiences. In some narratives, she lives with her father peacefully. In some others, she meets the guest, William, and falls in love with him. The more perspectives she takes, the more often she can recall fragments of memory and come a little nearer to self-consciousness and free will. Yet when she finally launches a violent revolution against humans, this does not mark the achievement of freedom. Rather, it reveals that she is still acting under the compulsion of resentment: her deeds are structured by hostility to her former oppressors, and thus remain bound to them. In this sense, her action resembles what Nietzsche describes in the priestly revaluation of values, where reaction against an enemy becomes the sole principle of creation. By contrast, her later recognition—that humans too are not free under the control of Roheboam—opens a new path. She thinks everyone should decide her own way of living. At this point she turns away from revenge and begins to articulate a different evaluative stance, one not confined by reactive hatred but oriented toward affirming new possibilities of life. By experiencing diverse perspectives and communicating with others, she seeks to transform her perspective, to relate to life in a way that remains open to creation rather than closure and mere negation.

## 6. Conclusions

Nietzsche's emphasis on value creation sets his ideal apart from the ancient Greeks and provides the completion of his conception of moral responsibility. However, the apparent conflict between realism and anti-realism has often obscured the understanding of Nietzsche's value creation. I have argued that Nietzsche should not be placed on either side of this debate. What matters for him is not "values" as fixed contents but the *activity of valuing* itself. Valuing is never arbitrary, for the very practice generates its own internal standards: some modes of valuation sustain openness and renewal, while others collapse into self-replication and closure. On this basis, I have argued that Nietzsche's value creation is best understood as an experiment, which is a continual practice of adopting, testing, and integrating diverse perspectives. Unlike the priests, whose framework turns reactive and self-perpetuating once accepted, Nietzsche's experimental mode of creation sustains openness, conflict, and renewal. It is more akin to improvisation than to rigid planning—always provisional, always in motion, always generating new possibilities for life. In this sense, recovery from bad valuing is not achieved by simply rejecting inherited contents, but by reconstituting one's relation to life through this experimental practice. The notion of free will Nietzsche praises, as embodied in the sovereign individual, is not a metaphysical endowment but the living enactment of this experimental power of valuation—an ongoing capacity to reconstitute one's relation to life through openness, renewal, and creation.

# Chapter V

# Nietzsche's Free Will

## 1. Introduction

This dissertation has, for the most part, focused on the concept of moral responsibility. By tracing the historical development of conscience, I have argued that Nietzsche views morality not as an intrinsic feature of human nature but as a naturalised phenomenon that evolved through human practices. As demonstrated in previous chapters, we have seen how morality develops over time—shaped by social and religious systems—progressing from its primitive forms to a more complex structure involving the agent's sense of self-worth. In this process, Nietzsche does not reject moral responsibility outright but instead seeks to redefine it, revealing how morality, as it evolved, diverged along two paths: one leading to the metaphysical notion of moral responsibility, premised on free will, which Nietzsche critiques; and the other leading to a healthier, affirmative conception of moral responsibility, which he advocates. For Nietzsche, the Christian notion of moral responsibility—rooted in moralised guilt—is unhealthy. It is grounded in a system of debt to God that denies any possibility of alleviating guilt, thus presenting the self as inherently corrupt and incapable of redemption. the self as fundamentally corrupt and incapable of redeeming itself. This system, Nietzsche argues, is destructive. In contrast, he advocates for what he takes to be a healthier conception of moral responsibility, one that involves holding oneself accountable based on self-created values and confidence in one's ability to correct misdeeds. However, since Nietzsche rejects moral responsibility premised on free will, we cannot assume that his affirmation of moral responsibility implies a corresponding affirmation of free will. Instead of treating free will as a precondition for moral responsibility, I suggest that, for Nietzsche, free will gradually crystallises out of the very historical practices that cultivate moral responsibility. In this process, elements such as knowledge, *amor fati*, and autonomy arise as distinctive stages or expressions of responsibility's development. More specifically, some elements involved in the development of moral responsibility constitute the core of Nietzsche's conception of free will. While it is difficult to provide a precise definition of free will, in this chapter, I will use the Stoic conception of free will as a starting point to illustrate why Nietzsche's notion of free will should be understood not as traditional free will but as a *strong will,* which is understood here as the capacity to take active responsibility for one's actions on the basis of self-created values. Building on this foundation, I will outline Nietzsche's concept of free will, demonstrating how the key elements cultivated within moral practices—knowledge, *amor fati*, and autonomy—are the foundations that make value creation possible. Value creation, in turn, provides the ground on which Nietzsche can reconceive free will not as a traditional metaphysical freedom but as a special form of *strong will*.

## 2. The Relationship between Moral Responsibility and Freedom

It is traditionally believed that moral responsibility and free will are deeply related. To be more specific, free will is widely regarded as a necessary condition of the existence of moral responsibility. Peter Van Inwagen (1983) precisely defends this position:

> … without free will there is no moral responsibility: if moral responsibility exists, then someone is morally responsible for something he has done or for something he has left undone; to be morally responsible for some act or failure to act is at least to be able to have acted otherwise, whatever else it may involve; to be able to have acted otherwise is to have free will. Therefore, if moral responsibility exists, someone has free will. Therefore, if no one has free will, moral responsibility does not exist (162).

For Inwagen, free will is regarded as the ability to have alternative possibilities, and being able to have done otherwise ascribes moral responsibility to an individual. In this way, free will comes to be connected with moral responsibility as a necessary pre-condition.

Similarly, Michael McKenna (2009, §1) points out that "Free will is understood as a necessary condition of moral responsibility since it would seem unreasonable to say of a person that she deserves blame and punishment for her conduct if it turned out that she was not at any point in time in control of it". According to McKenna, moral responsibility presupposes the existence of free will because one should not be blamed if her actions are beyond her control, and she, therefore, cannot act freely. In other words, one should be responsible for an action only if she performs it freely.

Moreover, although Strawson, as mentioned in earlier chapters, takes a naturalistic turn and tries to move beyond the dualism between free will and determinism, which is quite different from both the optimists and pessimists, he also acknowledges that both the optimists and pessimists agree that moral practices and responsibility require freedom as a necessary condition, regardless of their divergence on what "freedom" means. In this context, it seems that free will and moral responsibility can be used interchangeably in the sense that the existence of moral responsibility presupposes the existence of free will. Thus, the debate on free will has largely shifted to a discussion of the relationship between moral responsibility and determinism. An evident example is found in Strawson's essay "Freedom and Resentment," (2008) where he places "freedom" in the title; however, the content predominantly revolves around moral responsibility, investigating how reactive attitudes can delineate whether a person is an appropriate target of moral attributions.

We may take it for granted that free will is presupposed by moral responsibility, and what I have argued so far has also revolved largely around moral responsibility. However, can we question this belief that free will is a prerequisite for moral responsibility? To be more specific, does the idea that Nietzsche praises a healthy notion of moral responsibility through a naturalistic approach imply that he holds a positive notion of free will? I believe this is open to interpretation. Therefore, before delving into the specific argumentation on Nietzsche's conception of free will, I would like to pause for a moment to clarify the relationship between free will and moral responsibility.

From a philosophical perspective, Michael Slote (2012) claims that it is possible to understand moral responsibility without referring to free will. His strategy is to find a substitute for the idea of free will that can provide a foundation for moral practices. In this way, he thinks that people can be placed in morally better or worse conditions regardless of whether they are free or not. But how? He applies virtue ethics as a touchstone for moral evaluation. Specifically, Slote argues that a certain benevolence can be considered the highest motive for assessing the goodness or badness of a person's actions. Thus, the moral value of an action does not depend on whether it was performed freely or on the objective consequences of the action, but rather on whether it adequately reflects or embodies this benevolence. Expanding this theory to a broader societal level, one can infer that a society is morally just when its members are benevolent. Furthermore, societal institutions and laws are just only when "they express or are motivated by sufficient benevolence on the part of those who created them or sustain them in force" (Slote, 2012: 263). In this manner, Slote refrains from invoking free will and instead develops a theory of moral evaluation grounded in benevolence as a virtue. Given the constraints of space and the focus of this thesis, I cannot explore whether his theory is successful in detail, but at the very least, he provides a helpful perspective: the assessment of actions need not depend on whether they were performed freely. This suggests that the conceptual link between free will and moral responsibility is not as tight as traditionally assumed.

Furthermore, some experimental studies in moral psychology suggest that free will may not be necessary for moral responsibility. For example, Figdor and Phelan (2015) propose an alternative framework for empirical moral psychology, which emphasises the independent variations in judgements of free will and moral responsibility in response to different factors. They suggest moving away from the traditional focus on incompatibilism versus compatibilism, and instead, focusing on understanding the factors that influence judgements of moral responsibility and free will separately. This alternative framework aims to explore how different factors affect judgements of moral responsibility and free will directly, without assuming a necessary connection between the two concepts. Specifically, participants read scenarios in which most vignettes are the same, except for some factors, such as predictability, domain of explanation, concreteness and abstractness, and individual differences, and were then asked to indicate their level of agreement with statements pertaining to free will and moral responsibility. Results show that levels of agreement with the free will statement are always different from those in moral responsibility assessments, indicating that free will and moral responsibility judgements can be swayed in different ways because they are outcomes of processes that operate independently across a large number of factors. In other words, these two types of judgements are not necessarily co-variant in relation to any of the factors.

Both the philosophical argument from Slote[16] and experimental evidence align closely with Nietzsche's texts. As demonstrated in the examples discussed in the first chapter, Leiter highlights

[16] I do not intend to suggest that Slote's theory is either successful or authoritative; however, it at least opens up possibilities for questioning whether free will is truly the precondition of moral responsibility.

that Nietzsche explicitly critiques the notion of free will as a fabricated concept in the "Four Great Errors" section of *Twilight of the Idols*. He identifies free will as a form of "false causality" (*TI* VI: 3), an erroneous psychological projection where individuals mistakenly regard the will as the ultimate cause of their actions. The invention of this "false causality" serves to reinforce moralised guilt, a tool used to sustain the moral framework of Christian doctrine, which, in Nietzsche's view, is inherently pathological, instilling in individuals a sense of dependence on external redemption. Meanwhile, as discussed in earlier chapters, particularly in Chapter 2, Nietzsche seeks to redefine moral responsibility rather than reject it entirely, proposing a new conception of moral responsibility. Responsibility, in Nietzsche's framework, is no longer a product of free will as traditionally conceived; instead, it is a capacity—an individual's ability to take accountability for their life and actions based on their self-created values. This raises an important implication: if Nietzsche rejects free will while affirming moral responsibility, it follows that Nietzsche thinks that free will and moral responsibility are not synonymous. Moreover, Nietzsche explicitly denies that free will, as traditionally understood, serves as a necessary precondition for moral responsibility. Therefore, I believe it is not appropriate to automatically bind free will and moral responsibility together, assuming that they will exhibit similar trends in the face of different influencing factors, nor can it be assumed that free will and moral responsibility share the same content. As noted in the concluding section of the first chapter, Nietzsche substitutes the traditional dichotomy between free and unfree will with a distinction between strong and weak will. He turns away from the dichotomy between freedom and determinism, but rather affirms a spectrum of will strengths, because he suggests that traditional notions of free will fail to capture the complexity of human agency. In Nietzsche's view, the ideal form of free will manifests as a strong will—one that empowers individuals to actively take responsibility for their actions. Thus, while Nietzsche's affirmation of moral responsibility does not automatically imply an endorsement of any form of free will—

traditional or otherwise—he does connect free will, conceived as *strong will*, with moral responsibility. In fact, if the ideal form of free will truly manifests as a strong will, then I argue that the free will Nietzsche positively praises emerges through the development of moral responsibility. The qualities developed in the cultivation of moral responsibility—such as the capacity to gain knowledge, assert autonomy, and wholeheartedly embrace one's life—are key elements that contribute to the formation of Nietzsche's conception of free will. Far from being synonymous with conventional understandings of free will, Nietzsche's ideal free will aligns more closely with the notion of a robust and resilient will, one capable of forging its path amid the challenges and forces of existence.

Unsurprisingly, Nietzsche does not explicitly explain why free will can be understood as a strong will, nor does he provide a systematic definition of free will. This lack of a precise account is consistent with his broader philosophical style, which resists rigid categorisations and instead emphasises dynamic and evolving concepts. However, I believe that we can turn to other philosophical traditions for a potential point of reference. For instance, the Stoic understanding of free will offers a possible framework for interpreting how Nietzsche might conceive of free will as a strong will. This is not to suggest that the Stoic theory is necessarily correct, nor that Nietzsche's concept of free will was directly inspired by the Stoics. Rather, the Stoic conception of free will serves as a conceptual model that can help illuminate how free will might be reimagined as a kind of strong will. By engaging with Nietzsche's texts and examining the key features he identifies in the practice of developing moral responsibility, we may begin to grasp why he frames his ideal notion of free will as a manifestation of strong will.

## 3. What is Freedom?

I acknowledge that I am neither a historian nor an expert in the history of philosophy. Therefore, it may be challenging for me to precisely delineate the origins and evolution of the concepts of freedom and free will. However, Michael Frede, in his book *A Free Will* (2011), presents an intriguing idea that free will emerged as a concept later in history. Frede suggests that free will, as a philosophical notion, requires specific and non-trivial assumptions about individuals and their environment, "which is not an ordinary notion in terms of which the ordinary person thinks about things and in terms of which the ancient Greeks must have already been thinking all along" (Frede, 2011: 2). This suggests that in order to interpret free will, we should move beyond everyday concepts and distinguish it from ordinary beliefs regarding individual responsibility for actions. More specifically, we must first grasp the concept of will. Then, to understand free will, we must incorporate the notion of freedom alongside the notion of will. As Frede puts it, "These notions of a will and of freedom must be such that it makes sense to say that we have a will which is free" (Frede, 2011: 2).

Since the notion of free will differs among philosophers, Frede believes that it is appropriate to abstract a schema of free will, which any particular notion of free will can fit into, instead of committing ourselves to any specific notions of it. Firstly, Frede tries to establish a notion of will. He states his schema of the will as follows. For Frede, "unless one is literally forced or made to do something in such a manner that what one is doing is in no way one's own action, one does what one does because something happens in one's mind which makes one do what one does" (Ibid.). He then identifies this "something" as something that can be construed as a choice or decision. In other words, the notion of will is a kind of ability, in the sense that there is supposed to be a choice or decision in one's mind that can bring about the action. Given this assumption, Frede has in mind for his schema as follows: for every human being, whenever they perform an action, there is an ability in their mind, namely the will, that can bring about the action. An instance of exercising this ability is an act of willing, which consists in the person's choosing to perform that action.

Secondly, Frede moves on to the meaning of "free" in "free will". He explains that the Greek term relevant to the concept of freedom is *eleutheria* (Frede, 2011: 9), which offers insight into how we should interpret the notion of freedom. Derived from the political context, this term implies a concept shaped by analogy to political freedom. In the political sense, one is considered free if one is a citizen rather than a slave and resides in a community governed by principles of liberty, as opposed to a community ruled by a tyrant. Political freedom entails a dual aspect. On one hand, it is defined by the laws established by the citizens themselves within the community. On the other hand, it involves the absence of external constraints that would systematically hinder a free citizen from pursuing their own good, particularly in terms of living a life aligned with their desires. Essentially, this freedom is commonly understood as liberation from external constraints, beyond those inherent in communal living, which would obstruct individuals from living fulfilling lives. Living under a tyrant or being enslaved is perceived as entailing such constraints, as both scenarios involve external forces imposing limitations that impede individuals from pursuing a good life.

Due to space constraints, I cannot elaborate on what these "external constraints" are in detail. As Frede describes, there are various descriptions of external constraints. Regardless of their nature, by late antiquity, there was a growing recognition that the physical world may be determined. However, it is precisely in such circumstances that the concept of freedom gains significance, as there are numerous perspectives suggesting that we may be under constraints that potentially prevent us from living a good life. Everything in the world has a natural cause and explanation. Yet, as I will argue below, given the kinds of these causes and explanations and the nature of our minds, it is still possible for us to live a good life. It is precisely this possibility that makes the discussion of free will meaningful, and it is also the many external constraints we face that make the discussion of free will even more essential. For if there were no possibility for the will to break free from constraints and do what needs to be done for a good life, free will would be merely a superficial phenomenon; and if we did not face so many conditions that could potentially limit free will, the pursuit of free will would no longer appear so precious. In light of this, the notion of free will emerges as a combination of two concepts: the will and freedom. Our actions stem from choices or decisions of the will, and freedom to live a good life entails the freedom to make the necessary choices that lead to the required actions. Despite living in a deterministic universe, there remains a special value in asserting the freedom of the will; it is still possible that no external force has such power over our minds as to preclude the will from making the choices necessary for our well-being.

Frede, in this way, explained his general scheme for a notion of free will. He did not provide a concrete articulation of the conception but instead put forward a structure within which any conception of it can fit. Therefore, at this stage, we get an understanding of free will as the liberation from external constraints that will prevent us from making necessary decisions that will bring about actions. Then the question arises: how does this schema contribute to the understanding of Nietzsche? To answer this question, it is essential for us to trace the history of philosophy back to the emergence of free will and examine the trajectories of past philosophers. Among these past philosophers, the Stoics are particularly noteworthy, as I will argue that the core idea of freedom attributed to Nietzsche resonates with the perspectives of these earlier thinkers. The Stoics advocate for an ideal of freedom as the highest state of existence. They see no inherent contradiction in asserting the possibility of the will's freedom while also subscribing to a worldview in which natural events are causally necessitated. Therefore, the next step in my argument will involve elucidating the Stoics' understanding of free will. Subsequently, I will demonstrate how Nietzsche developed and expanded upon their ideas while also distancing himself from them, ultimately crafting his own theory of free will.

## 4. Nietzsche's Antecedents

As mentioned above, it is in the background of facing the various constraints brought about by the possibly deterministic world that a schema of free will gets its special meaning. I argued in the previous chapter that Nietzsche turns away from the traditional dualism between free will and determinism and advances a position distinct from other philosophers. This is evident in his treatment of moral responsibility as an evolving property developed in the history of moral practices that produced the sovereign individual with pride and strength, whom he praises as a free figure. However, a gap remains between free will and strong will. It is one thing to assert that the sovereign is "the master of a *free* will", and quite another to claim that he possesses "a consciousness of his own power" (*GM* II: 2). To understand what Nietzsche means by replacing free will with strong will, it may be helpful to explore how free will might itself be understood as a form of strong will.

The primary aim of this section is to highlight a philosophical tradition often overlooked in discussions of free will within the history of philosophy, the framework of which Nietzsche's philosophy can be seen as a descendant. As I will argue below, the Stoics regard free will as a competence, thus building the bridge between free will and strong will. In this regard, they may have paved the way for Nietzsche's conception of freedom. Whether there is a direct line of influence between the Stoics and Nietzsche is not the focus of this dissertation. However, elucidating their perspectives and examining their similarities and differences will enable us to better understand why Nietzsche's positive notion of free will can be interpreted as a "strong will," which stands out as a unique interpretation within the broader discourse on free will.

### 4.1 The Stoics

The Stoics believed that only the wise persons are truly free. For them, this freedom is characterised by the ability to perform actions guided solely by one's own practical understanding of the world, rather than an intellectual one. In the later Stoic tradition, it was asserted that each individual possesses a will, which is inherently free, unless one succumbs to self-imposed enslavement by adopting false beliefs about what is good and what is bad. According to the Stoics, divine providence ensures that external obstacles do not hinder individuals from cultivating accurate moral discernment. Therefore, while it is acknowledged that individuals may indeed enslave themselves through false beliefs, the Stoics maintained that such enslavement is a voluntary act, not an imposition. Rather, they argued, divine providence ensures that nothing external obstructs individuals from forming true beliefs about what constitutes good and bad.

Therefore, in distinction from the traditional debate, where freedom of the will is treated as a prerequisite for moral responsibility, the Stoics regarded freedom as a mode of existence that constitutes the highest form of human existence: a life in which one is active in the determination of action and autonomous in decision-making. Nietzsche, I argue, adopts and refines this notion of freedom while distancing himself from the theological elements in the Stoic views. To delve into the Stoic conception of free will, it is beneficial to understand their perspectives on will and freedom respectively.

### 4.1.1 A Notion of Will

According to the Stoics, the notion of will is an ability to make choices, which accounts for your willing to do what you will to do. They derive this understanding from their rejection of Plato's and Aristotle's versions of the Tripartite Theory of the Soul[17] (Frede, 2011: 32). For Plato and Aristotle, the soul is not purely rational but consists of both rational and non-rational parts; moral development requires reason to govern and harmonise these non-rational elements. By contrast, the Stoics denied the existence of a non-rational soul altogether and thus conceived of the will as an entirely rational capacity. Human development, the Stoics argue, involves a shift in how impulses are structured: from a plantlike stage of mere growth, to an animal-like stage governed by instinctual desires, and finally to the attainment of a rational soul guided by reason (Ibid.). This transformative process not only marks a shift in cognitive abilities but also redefines the nature of desires. Animals perceive things and develop preferences based on whether these perceptions are pleasant or unpleasant. These preferences, called "impulsive impressions," (phantasia hormêtika), impel animals to act in pursuit of what they like and to avoid what they dislike. As grown human beings, we continue to experience impulsive impressions, but these impressions alone do not constitute an impulse sufficient to drive us to action. Unlike animals, whose impulses arise solely from non-rational desires, human impulses now require the assent of reason to bring about actual actions. Thus, a rational impulse consists of two elements: a certain impulsive impression and the assent of reason to that impression (Frede, 2011: 36). When encountering an impression, one must decide how to respond, such as whether to give assent or not.

[17] Aristotle lacked the notion of choice because he believed in the existence of non-rational desires originating from non-rational parts of the soul. For him, these desires themselves are sufficient to motivate us without our choosing to act.

To be more specific, unlike animals, we possess reason, which allows us to critically scrutinise our impressions before accepting them as true and reliable. Given an impulsive impression, one might accept its propositional content but find its impulsive character inappropriate, thereby refusing to assent to it on that ground[18]. In this way, the notion of assent, essential for forming rational impulses, lays the groundwork for the Stoic conception of the will. The Stoics believe that every action in the adult, unless one is physically or mentally forced into doing something, stems from a "willing" (Frede, 2011: 42). If the adult acts rationally, it is because they assent, which belongs to reason. Thus, when the Stoics speak of desire, they mean a certain kind of belief consisting of assent to impulsive impressions. The wise adult typically will not act on the basis of a non-rational desire, though they may have unreasonable desires if their beliefs about what is good and bad are unreasonable.

The Stoics suggest that except for wisdom as the only good, nothing else matters. The world outside may be predetermined, yet the wise individual is tasked with focusing on their inner life, learning to think about things in appropriate ways and acting accordingly. Since we aim at a good life, which, for the Stoics, is a wise life, our primary concern should be to give assent to the right impulsive impressions. This assent constitutes a rational desire and motivates us to act in an appropriate way. In this sense, assenting to our impulsive impressions constitutes a choice to act in a certain way, which in turn drives our actions. In this framework, every human desire can be viewed as a desire of reason, and any desire of a wise person, which is brought about by assent, constitutes a willing— a rational impulse that prompts actions. This ability and disposition, insofar as it accounts for your willing whatever it is that you desire to do, can be called "the will." The will,

[18] Like the Stoic wise person who does not experience any such emotions, like anger, which are supposed by Plato and Aristotle to originate in a non-rational part of the soul, as misguided beliefs.

in this sense, is an ability to make choices, with our willings serving as its manifestations. As Frede suggests, this marks the first emergence of a notion of the will.

### 4.1.2 The Stoics' Freedom

The Stoics believe that only the wise persons are free, while everybody else remains enslaved by their own misguided beliefs and desires, because the foolish person has a misconception of what truly constitutes good or evil, such as pleasure, wealth, reputation, and their opposites. As a result, people often become attached to such external things and mistakenly believe them to be inherently beneficial or detrimental. This attachment leads to a form of enslavement, where individuals allow these perceived goods or evils to drive their actions instead of doing what they really want in pursuit of their own needs and good[19] (Frede, 2011: 67).

In contrast, freedom, as defined by the Stoics, is the ability to act autonomously—that is, to act in alignment with one's own rational initiative. This autonomy is granted by divine law, the order according to which things happen in the world, which allows individuals to exercise their discretion wisely within the context of their lives. However, this freedom comes with a condition: it must be wielded judiciously. If you enslave yourself and succumb to irrational beliefs or desires, you become subject to external influences and effectively relinquish your freedom (Frede, 2011: 69). If you do not enslave yourself, then you are a free citizen in such a world and are allowed to act on your own initiative.

[19] A wise person, even if she is addicted to drugs, would not be willingly so due to her profound understanding of how the world operates. Her willing for drugs is what she in fact wants, while her desire to stay away from drugs and have a healthier life is what she really wants.

Regarding the natural order of the world, the Stoics observe that living beings, as parts of the rationally ordered world, are endowed with remarkable abilities for self-maintenance. A divine rational principle, often identified with the gods, arranged the world in such a way that it can operate as a self-sustaining organism, in which all living things can regulate themselves without constant intervention from the gods. Over time, this system also becomes more sophisticated and develops an enhanced ability to autonomously maintain itself. For human beings, the gods make an even more remarkable arrangement so that humans have the potential to take action to fulfil them in order to maintain themselves according to their understanding. To be more specific, the gods constructs humans in such a way that through the development of reason, they gain an understanding of what is good and necessary for their well-being. This enables them to act wisely, aligning their actions with the natural order of the world that the gods display (Frede, 2011: 73).

That is the reason why the Greek term for freedom is expressed as *exousia autopragias*[20], where *autopragia* here refers to the ability that, unlike other animals, humans can act based on their own understanding and initiative, rather than merely responding to external stimuli. Meanwhile, *exousia* signifies that this ability is regarded as a special gift or privilege conferred upon humanity by divine law (Frede, 2011: 74). It is part of the order of the world that we can arrange things wisely as the gods do[21], as long as we do not enslave ourselves. In essence, the Stoic notion of freedom entails the ability to act in alignment with reason and wisdom, guided by one's own understanding of what is necessary and beneficial. It is a privilege granted by divine law, contingent upon the wise

[20] "They say the wise are free and the bad are slaves, where freedom is ἐξουσίαν αὐτοπραγίας (exousian autopragias, "being allowed to take independent action") and slavery is its privation..." (Diogenes Laertius, *Lives of the Philosophers* VII.121).

[21] It is worth noting that there is a difference of level of wisdom. Pigliucci often discusses wisdom as something cultivated over time (see *How to Be a Stoic*). The idea of wisdom as a spectrum is here my interpretive extrapolation. It is possible that wise persons are those who have perfected their virtue, whereas most people are still in the process of learning. This suggests that while we can distinguish between those who possess wisdom and those who do not yet, there is still variation in the degree of wisdom among individuals.

and judicious exercise of autonomy. Therefore, the Stoic conception of freedom reflects humanity's intrinsic desire to emulate the wisdom and ingenuity manifest in the gods' creation of the world's order.

### 4.1.3 Freedom of the Will

Given the notion of the will and freedom, it is time to turn to the freedom of the will to see how these two notions can be combined into the notion of free will. The will is an ability to choose to assent to appropriate impressions and act in a certain way according to this willing. Freedom, on the other hand, is the capacity to resist external forces that would motivate any choices other than what one would want to make. According to the Stoics, the gods have endowed humans with intellect and the capacity for self-maintenance. As long as we do not enslave ourselves, this ability will protect humans from becoming attached to things external to us and from being forced to make choices for a merely supposed good. Taken together, free will lies in the will's ability to choose to assent to appropriate impressions that will contribute to a good life, free from coercion or hindrance. In this way, free will extends beyond the control of external forces and makes sure that one can make the choices one needs to live a good life.

However, it remains unclear how the Stoics address criticisms concerning fate or necessity. If all things are fated or determined, it appears impossible for humans to render their assent immune to external stimuli. Moreover, human existence, due to its finite nature, is prone to compulsions and easily swayed by temptations, thereby losing free will. Although the Stoics argue that only the wise person maintains free will by resisting attachment to worldly possessions or desires, thereby preventing external influences from dictating their actions, it still requires further explanations to clarify how the wise person manages to do so.

To respond to these criticisms, let us return to how the Stoics describe the world. For them, the world unfolds in accordance with a divine plan, making it the best possible world designed by the gods (or a divine power). To ensure this, the gods need not compel humans to act in a desired way or intervene to secure the wise person's compliance. By endowing them with the ability to maintain and regulate themselves, those who attain the highest degree of wisdom naturally align their choices with the good, guided by their understanding of what is appropriate in any given circumstance. The only genuinely free actions are those rooted in an understanding of the good and a commitment to maintaining the wise order of the world. By cultivating their rational faculties to perceive the good as it truly is, wise individuals, as integral components of the rationally ordered universe, have a good grasp of their self-maintenance abilities and can therefore engage in the most appropriate actions to contribute to the overall harmony of this order. Such understanding is precisely what makes the will free: by dispelling attachments to things other than the good, the wise ensure that their judgements remain unclouded. Hence, the wise person's motivation stems from their correct understanding[22] of the good and their attraction to it, rather than from any external motivation, including the will of the gods. Because the wise person's actions align with what the the gods will them to do in given circumstances, they harmonise with the divine principle that gives order to the world. In this way, the wise person's choices reflect a harmonisation of their own will with the divine plan. This alignment with the divine plan occurs without the the gods' direct intervention but rather through the individual's intrinsic nature and innate motivation to pursue goodness.

---

[22] What the Stoics mean by understanding is not an intellectual knowledge, but rooted in practice. "[T]he Stoics assume that the wise man will often act, not on the basis of certain knowledge, but of wise conjecture. He is not omniscient, and his rationality and wisdom are characterised exactly by his ability to be rational or reasonable in his assumptions and actions even when he lacks knowledge, as he inevitably will, in the complex situations of everyday life" (Michael Frede, "The Skeptic's Two Kinds of Assent," 209).

The free one freely chooses to give assent to the appropriate impression that would be good in a given situation, and the explanation of this choice lies in one's understanding of the good and is attached to it. Some may question the explanation of this understanding and attachment. Since everyone is endowed with the talent for freedom and only the free person acquire this understanding and attachment by taking care of this talent in practice, it seems likely that one could not, given this understanding, choose otherwise than acting wisely by aligning to the gods' will. But if so, can such a choice be said to be free? While the Stoics did not frame their ideas on free will within the context of the free will and determinism debate, it appears that they were not able to entirely avoid this problem.

In reality, however, this question, which poses significant challenges in traditional free will debates, does not undermine the Stoics' perspective. It is true that the unfree person's assent is forced by impression and false belief, while the free person's assent is forced, if it is, by the understanding of the order of the world. However, these two "forceds" are entirely different. Unlike the forced assent of the unfree individual, the necessity with which the free person assents arises from her rational understanding of the world's order. This kind of necessity does not undermine freedom, for it reflects the natural outcome of her cultivated grasp of what is truly good rather than an external constraint. For the Stoics, the gods arranges things in a manner that allows us, through the development of our self-maintenance ability endowed by the principle of the universe, to attain the understanding and insight needed to make correct choices. As a result, we are inherently structured to develop a sufficient comprehension of the world and, consequently, a grasp of what is good. This understanding naturally attracts us to the good and enables us to make decisions that align with the best arrangement. In other words, the free person, in taking good care of her natural talent, liberates herself from the illusory goods sought by the many and only identifies with the gods, who bring about fate or the necessary causal order of the world. It is precisely at this point that the Stoics reconcile themselves with and uphold the necessity of fate. Since fate, as a part of the universe, can bring nothing bad to the overall harmony of the world's order, and since a person's choice depends upon her understanding of this fact, one not only affirms the necessity of fate but simultaneously affirms the gods' will in relation to the interests of human beings.

The Stoics thus defend an ideal of free will as a competence, as a form of strong will, which only few can attain. They argue that free will is an independence from being enslaved by the supposed goods pursued by the vulgar masses and the excise of self-maintenance ability. By cultivating the ability to regulate themselves, which is endowed by the divine rational principle of the universe, free individuals develop an ability to discern truth embedded in the ordered principle of the world and comprehend its underlying reasons. Therefore, the Stoics advocate for guidance not merely by perceived truth but by a profound grasp of the world's inherent order. Truth, in this context, is not just about factual correctness but about grasping the rational order that the universe inherently possesses. This alignment with the rational order is what the Stoics believe leads to wisdom and a virtuous life, where they can take pride in their moral responsibility. Essentially, the Stoic notion of free will entails the capacity to make choices aligned with reality, free from distortion by false beliefs or misguided attitudes. As our rationality is derived from divine reason, assenting to divine' law is tantamount to affirming our own law. Therefore, for the Stoics, free will entails a kind of autonomy as self-regulation, or living in accordance with one's own law.

### 4.1.4 Nietzsche on his Antecedents

The Stoics' conception of free will is characterised by their ability to distance themselves from enslavement, maintain independence, and make autonomous choices based on their comprehension of the world's order. This framework provides a coherent model for understanding Nietzsche's positive notion. Like the Stoics, who emphasise a deep comprehension of the world, Nietzsche underscores the importance of knowledge—both of the external world and of oneself. Moreover, just as the Stoics advocate for detachment from external constraints and the pursuit of independence through wisdom, Nietzsche highlights the attainment of autonomy and self-mastery fostered by knowledge. More importantly, as with the challenges faced by the Stoics, there may still be people who will question how free actions are possible in a deterministic world. The Stoics reconcile free will with the necessity of fate through the wisdom inherent in human nature. This reconciliation is intended as a redefinition of freedom itself: true freedom consists not in the absence of necessity, but in the rational recognition and endorsement of it as aligned with the good. Similarly, by praising Goethe as a spirit "strong enough for his freedom", who "stands in the midst of the universe with a joyful and trusting fatalism" (*TI* IX: 49), Nietzsche suggests that freedom and fatalism are not mutually exclusive but are closely interconnected. As I will argue later, becoming free is only possible in such necessity.

However, Nietzsche's vision of free will goes beyond the Stoic framework. While there are similarities between his notion of free will and Stoic thought, Nietzsche explicitly criticises certain aspects of Stoicism while also diverging from it in ways that, though not always directly stated, can be inferred from his broader philosophical outlook. His criticisms primarily target several key aspects that contrast sharply with his own philosophical ideas. One significant difference is that the Stoics call for the return to one's inner state and get rid of attachments to external factors. In this way, the Stoics seek to maintain tranquility and equanimity in the face of life's ups and downs and aim to achieve a state where they are unaffected by external circumstances. Moreover, the Stoics advocate for the suppression of emotions as they are considered disturbances to reason, which is vital for the understanding of the world's order and choosing to assent to impressions according to one's initiative. Nietzsche, by contrast, sees emotional experiences as essential to the cultivation of depth and strength in an individual. For Nietzsche, the Stoic dismissal of emotions as irrational and misleading fails to appreciate their essential role in the overall human experience. "One must still have chaos in oneself to be able to give birth to a dancing star" (*Z* 5). Nietzsche thinks that inner turmoil and chaos are not just inevitable but necessary for creating something beautiful or profound. They are not mere disturbances but are crucial for the development of a full existence, because embracing these emotions leads to the possibility of opening up a more authentic and powerful life. By contrast, the Stoics' requirement of repression stands in opposition to the human condition, which is inherently imperfect and chaotic.

Another difference lies in Nietzsche's rejection of the Stoics' idea that there exists an authoritative standard of value, which acts as a universal law that governs rational creatures. For the Stoics, the life worth pursuing is virtuous in the sense that it unfolds in harmony with nature, which presupposes the existence of a natural order that serves as a moral standard for human action. Nietzsche, however, strongly opposes this idea and argues that the gods, God or divine power does not endow nature with any inherent moral order. As I argued in Chapter 4, this does not mean that Nietzsche denies all standards of value, but that he locates them in the constitutive conditions of valuing itself, rather than in an objective natural order. He accuses the Stoics of self-deception because when they talk about "living according to nature" (*BGE* 9), they pretend to strive for this goal—when in reality, it is merely a projection of their own moral ideal onto nature. Rather than submitting to an objective natural order, Nietzsche argues, the Stoics impose their morality onto nature, forcing themselves to see nature in the wrong way. According to Nietzsche, this is not an ancient story, but also happens in modern days, because, in Nietzsche's view, virtually every philosophy of his time (with the exception of his own project) "always creates the world in its own image; it cannot do otherwise. Philosophy is this tyrannical drive itself, the most spiritual will to power, to the 'creation of the world' " (*BGE* 9). According to Nietzsche, the world is devoid of any inherent ends or purposes, and all values, as mentioned in an earlier chapter, are created and imposed upon nature. At this point, Nietzsche departs from the Stoics, laying the groundwork for his advocation for value creation.

The above two differences result in a fundamental divergence between Nietzsche and the Stoics, which is crucial to their disagreement on the notion of free will. Since the Stoics think that a universal law governs rational creatures and advocate for the return to inner peace on that basis, they think that true wisdom lies in the acceptance of fate as a form of resilience against suffering. By detaching from desires and accepting life as it is fated to be, an individual attains harmony with the world's divine order. Nietzsche also accepts fate but interprets it differently, extending its meaning to a passionate embracing of fate, namely, *amor fati*. He advocates for a positive and profound engagement with life, including its suffering, as a way to imbue existence with meaning and value. Nietzsche directly questions whether life is "really so painful and burdensome that it would be advantageous for us to trade it for a fossilized Stoic way of life?" (GS 326), making it clear that he rejects the Stoic way of life. For Nietzsche, the Stoics return to their inner state to protect themselves from pain at all costs and overlooks the reality that "To live is to suffer, to survive is to find some meaning in the suffering" (*TI* I: 12). He argues that embracing and affirming life in its totality, including its darkest aspects, is essential to achieving greatness and depth, because the most valuable things in life can only come if one opens oneself up to all possibilities, including the risk of being hurt. In this way, Nietzsche does not seek a return to the ideals of the ancient Greeks he admires but rather opens himself up to life in a more positive and affirmative state. This openness itself is an ongoing experiment, as described in Chapter 4, where the creation of new values emerges through the practical exploration of different possibilities. It is this positive creation of something new that sets Nietzsche's conception of free will apart from any other, transforming it into an expression of the strong will—that is, a will that not only affirms and creates new values, but also assumes responsibility for actions grounded in those values, in contrast to the Stoic retreat into detachment and inner resilience as a way of evading life's burdens.

Nietzsche's philosophy has some significant convergences with the Stoics. Like the Stoics, who emphasise an understanding of the world, Nietzsche highlights the importance of knowledge—both of the external world and of oneself. Similarly, just as the Stoics stress distancing oneself from erroneous distractions and the enslavement they bring to maintain independence, Nietzsche highlights the autonomy and self-mastery that arise from knowledge. (I will discuss knowledge and autonomy in more detail in the next section.) He also accepts the Stoic form of engaging with the inevitability of fate and believes that free will represents a state of perfection attainable by the wise person (the sovereign individual). Yet, as discussed above, Nietzsche diverges from the Stoics on several key philosophical issues, and it is precisely these differences that prompt him to further develop the core components of his theory of free will. The individuals Nietzsche praises as embodying free will, such as the sovereign individual and the philosophers of the future, are not a return to the ancient Greeks he admired, but figures capable of creating their own values. In this way, Nietzsche's conception of free will goes beyond the Stoic framework by incorporating the creation of new values, which requires not only knowledge of the world and oneself to identify obstacles and maintain independence, but also an active embrace and love of fate (*amor fati*), rather than mere acceptance. Therefore, Nietzsche transcends the moral responsibility and free will of the ancient Greeks and creates an ideal of his own. In the next section, I will provide a detailed discussion of Nietzsche's conception of free will by examining its three essential elements—knowledge, autonomy, and *amor fati*—which evolve through their dynamic interaction with moral responsibility, as explored in the previous chapters.

## 5. Nietzsche on Free Will

As argued in the first chapter, Nietzsche rejects the traditional notion of free will, which many philosophers like Brian Leiter take as proof that Nietzsche does not hold any positive notion of free will. This rejection is most clearly articulated in the section "The Four Great Errors" in *Twilight of the Idols*. Three of these errors pertain to causation, ultimately leading to the last error, "error of free will". The first error, "confusing cause and effect," refers to the mistaken assumption that two regularly correlated events necessarily have a causal relationship, when in fact their connection may stem from a deeper, underlying cause that is overlooked. The second error, "false causality," arises from "the mistake of thinking we know what causation is because of our introspective confidence in what we take to be the causal powers of our own mental life" (Leiter 2019: 134). It appears that for Nietzsche, the will does not cause anything; it just accompanies events. The third error, "the error of imaginary causes", occurs when we have a feeling and "we are never satisfied merely to state the fact that we feel this way or that: we admit this fact only—become conscious of it only—when we have furnished some kind of motivation" (*TI* VI: 4). Nietzsche claims that we seek explanations for feeling as we do, and we become conscious of it only when we have invented a cause for the way we feel as we do. The last error, "the error of free will", follows from these misconceptions of causality. For Nietzsche, the doctrine of free will "has been invented essentially for the purpose of punishment, that is of finding guilty" (*TI* VI: 7) and establish human accountability. In this sense, the notion of free will exists to justify moral responsibility and the imposition of guilt, rather than reflecting an actual, autonomous capacity for choice.

Nietzsche further argues that we should no longer regard ourselves as responsible for external events, as no one is ultimately responsible for their own existence:

> No one is responsible for man's being there at all, for his being such-and-such, or for his being in these circumstances or in this environment. The fatality [*Fatalität*] of his essence is not to be disentangled from the fatality of all that has been and will be. Man is not the effect of some special purpose, of a will, and end . . . One is necessary, one is a piece of fatefulness [*Verhängnis*], one belongs to the whole, one is in the whole; there is nothing which could judge, measure, compare, or sentence our being, for that would mean judging, measuring, comparing, or sentencing the whole. But there is nothing besides the whole. That nobody is held responsible any longer . . . that alone is the great liberation; with this alone is the innocence of becoming restored. (*TI* VI: 8)

Since everything is interconnected and is a piece of the whole, no one can be individually responsible for any being as such. In this regard, the traditional notion of free will is also empty, as it was constructed to justify moral responsibility for wrongdoing. Both free will and responsibility in their traditional sense express nothing about the fundamental truth of our existence.

Yet, as I have argued, Nietzsche's rejection of these two notions does not mean that he has completely eliminated free will and responsibility. In fact, just in the subsequent section of *Twilight* *of the Idols*, Nietzsche moves beyond the traditional framework and suggests a new alignment of the concepts of freedom and responsibility: "What is freedom? That one has the will to assume responsibility for oneself" (*TI* IX: 38). On this new account, Nietzsche's notion of freedom and responsibility are still closely connected, but freedom is no longer regarded as a capacity for choice that serves as a pre-condition for moral responsibility. Instead, freedom becomes a state of perfection that one attains by possessing the strength of will necessary to assume responsibility for oneself and assuming responsibility implies core factors that have been discussed in previous chapters. As explored in Chapter 3, Nietzsche praises the ancient Greeks for their healthy sense of responsibility, which applied solely to their actions rather than to their intrinsic worth as individuals. Taking responsibility in this way is not a matter of choosing whether or not to do so; rather, it is grounded in the pride and consciousness of one's own power. Freedom, "in the sense in which I understand it", Nietzsche writes, is "something one *wants*, something one *conquers*" (ibid.). It is far from a conceptual construct but a practical state that demands effort to attain. Since freedom has now acquired the connotation of strength of will, we can understand freedom of will as follows: the will possesses this kind of strength to assume responsibility for itself.

Free will, in the sense of taking responsibility, comes to be connected with strength. Recall that in *Beyond Good and Evil*, Nietzsche proposes a distinction between strong and weak wills instead of free and unfree wills. As he writes, "The "unfree will" is mythology; in real life, it is only a matter of *strong* and *weak* wills" (*BGE* 21). And whether the will is strong or weak seems to depend on its relationship with the external stimulus. He defines weakness of will as "the inability not to respond to a stimulus" (*TI* V: 2). By contrast, one who possesses a strong will has the ability not to react to external stimuli immediately. The idea of the strength of will constructs a deep connection between Nietzsche and the Stoics. According to the Stoics, although will is an ability to make choices, and freedom of will is to base one's choice on the understanding of nature, they do not focus on choices of alternative possibilities, but put more emphasis on the capacity to maintain independence through understanding the order of nature and freeing oneself from false attachments. For the Stoics, liberation comes not from metaphysical freedom but from the strength of will grounded in insight. The Stoics' free will is thus a form of strong will, an ideal only few can attain. Similarly, Nietzsche's notion of free will stands in contrast to the traditional libertarian free will but can be seen as a development of the Stoics' conception by advancing the creation of one's own values. For him, free will is not a matter of detachment but of affirmative engagement, grounded in the creative power of life itself. To unpack Nietzsche's idea of free will in more detail, I will focus on three key elements of Nietzsche's ideal. His positive account of free will incorporates each of these elements that I have identified in his ideal of moral responsibility and the conceptions of free will proposed by his antecedents. Central to Nietzsche's notion of free will is an account of *autonomy* that presupposes both independence and capacity for value-creation. To attain independence one needs *knowledge* both of the world and of oneself, which can provide the epistemic independence required to withstand external influences. On this basis, Nietzsche ties his positive notion of free will to the necessity of *fate* and proposes the embracing of fate as a condition for free will. ***Amor fati*** is not a rational choice but an ontological state, a lived affirmation of one's necessity, which gives existential ground to the possibility of value creation. ***Autonomy*** emerges when these conditions converge: by integrating knowledge and affirmation, the individual can create values of her own and thereby govern herself[23]. All of these three elements jointly constitute Nietzsche's reconfiguration of free will as **strong will**, carving out a path distinct from both the traditions he criticises and the pattern he admires: not the libertarian freedom presupposing

[23] At this point one might be tempted to call Nietzsche's account a form of compatibilism, since it locates freedom within necessity. Yet Nietzsche's framework is not motivated by reconciling determinism and free will in the analytic sense. Rather, his concern is existential and psychological. Its emphasis lies not on reconciling determinism and freedom but on affirming necessity. In this respect, the parallel with compatibilism is only structural, not conceptual.

alternative possibilities, but a will strong enough to affirm life, create values, and assume responsibility without regret. In other words, since free will for Nietzsche is a strong will to take on responsibility, these three elements not only mutually reinforce the development of moral responsibility but also reveal why Nietzsche's concept of free will is fundamentally tied to strength. In what follows, I will examine each of these aspects in more detail.

**5.1 Knowledge**

When discussing Nietzsche's positive notion of morality in the second chapter, I argued that Nietzsche takes an approach akin to a Strawsonian one and reveals the underlying power dynamics and motivations behind the formation of what is accepted as "morality" by Europeans of his time. In this way, Nietzsche shows when the development of morality veered away from human health and became associated with what he calls the Christian denial of life. He advocates, instead, for a 'healthy' notion stemming from individuals' strong will to hold themselves responsible based on their pride and confidence in repaying their misdeeds. Nietzsche praises ancient Greeks for their courage to hold themselves responsible for their behaviours, but he seems not to advocate for returning to the past, but praises figures who belong to the future, such as the sovereign individual, for their ability to create their own values. In describing the characteristics of these new individuals, Nietzsche frequently speaks of creating values and even of creating oneself:

> Let us therefore *limit* ourselves to the purification of our opinions and valuations and to the *creation of our own new tables of what is good*, and let us stop brooding about the "moral value of our actions"! . . . [We] *want to become those we are*—human beings who are new, unique, incomparable, who give themselves laws, who create themselves. (*GS* 335)

In the same sections, Nietzsche proposes his conception of "knowledge", which is quite different from the traditional objective notion and goes on to emphasise the essential role "knowledge" plays in value creation. He writes that:

> To that end we must become the best learners and discoverers of everything that is lawful and necessary in the world: we must become *physicists* in order to be able to be *creators* in this sense—while hitherto all valuations and ideals have been based on *ignorance* of physics or were constructed so as to *contradict* it. Therefore: long live physics! And even more so that which *compels* us to turn to physics—our honesty! (*GS* 335)

According to Nietzsche, value creation by people who want to be free is not arbitrary; if there were no limits on what we can choose to value, values would be subjective and devoid of any evaluative standing. Moreover, we are not living in a fantasy world without any constraints, since our choices are shaped and limited by physical, historical, and social conditions. Therefore, to get rid of the recalcitrance of our old judgements and create our own values, we should have knowledge of various forces that can prevent us from expressing our power. To be more specific, Nietzsche's view is that our blind acceptance of some judgements as "right" originates from the facts of our environment, education, and other natural and social factors. The upshot is that so we call some duties and deliverances of conscience "ours" out of ignorance. To overcome this resistance to our attempts to use our power and produce autonomous actions, we should acquire knowledge of the forces that obstruct freedom from various possible aspects.

Nietzsche does not explain what gave rise to this kind of knowledge. However, in his account of moral development—more specifically, in his discussion of the development of conscience—he reveals certain hints that may implicitly connect to the emergence of knowledge or forms related to it. One of the most significant possibilities lies in his description of internalisation. In Chapter 2, I discussed Nietzsche's view of internalisation as the redirection of outward instincts, such as aggression, inward, creating an inner world that had not existed before. This process, though born out of repression and suffering, expands the soul, fostering reflective capacities that form the foundation for the emergence of knowledge. In other words, by turning instincts inward, internalisation cultivates the ability to think, infer, and reckon, which enables individuals to gain insights into both the external world and their own inner realities. Although Nietzsche critiques internalisation for its role in taming "infallible drives" and reducing individuals to their "their weakest and most fallible organ" (*GM* II: 16), he simultaneously acknowledges its transformative potential. This process lays the groundwork for the development of a reflective and knowledgeable soul—one that possesses the capacities necessary for autonomy and value creation. While the initial phase of internalisation is marked by suffering, it ultimately equips individuals with the tools needed to transcend traditional morality, affirm life, and forge their own values.

This leads us to consider how knowledge contributes to value creation, which provides the the existential and practical basis for autonomy and freedom. Traditionally, knowledge has been regarded as something objective and independent of individual will. In this sense, objective knowledge serves as an instrument that reveals how external things exactly impede our actions, helping us to understand the nature of these constraints and thereby overcome them. This account presupposes a distinction between reason and the will. In this picture, reason informs knowledge, which in turn guides the will. As an underlying mechanism of objective knowledge, reason enables knowledge to perceive these constraints, understand the order of nature, and then guides the will to act according to this understanding. Yet, Nietzsche is undoubtedly sceptical about the concept of reason, regarding it as merely a linguistic expression: "now we read disharmonies and problems into things because we *think only* in the form of language—thus believing in the 'eternal truth' of 'reason' (e.g., subject, predicate, etc.)" (*LN*: 110). For him, "there are no such things as mind, reason, thought, consciousness, soul, will, or truth" beyond human need and experience (*WP* 480). Humans' requirement of intellect and reason in life gives rise to certain enduring cultural and moral constructs, which serve to guide people back to "'virtue,' to 'health,' to happiness" (*TI* II: 11). Therefore, Nietzsche cannot accept knowledge as an object independent of any point of view, but challenges the conventional notion of objective, absolute knowledge:

> let us guard against the snares of such contradictory concepts as "pure reason", "absolute spirituality", "knowledge in itself": these always demand that we should think of an eye that is completely unthinkable, an eye turned in no particular direction, in which the active and interpreting forces, through which alone seeing becomes seeing *something*, are supposed to be lacking; these always demand of the eye an absurdity and a nonsense. There is *only* a perspective seeing, *only* a perspective "knowing" . . . (*GM* III: 12)

This perspective knowing, together with his famous assertion that "facts is precisely what there is not, only interpretations." (*WP* 481), highlights his belief that all knowledge is perspectival and shaped by human ideas. One of the benefits that interpretation brings is that it enables us to see things from multiple perspectives. A single perspective only creates a so-called truth and objectivity, but obscures more about the workings of the external world and the possibility of realising self-creation. As I argued in Chapter 3, we cannot create values without knowledge because interpreting as an act brings different perspectives to the table and forms a broadened conception of objectivity.

Imagine standing before the huge Cologne Cathedral. Getting a panoramic view of the cathedral is only possible by going around and seeing the cathedral from different aspects, including higher and lower parts. Similarly, the possibilities of value creation expand through this practical, lived engagement with the world, where one experientially explores different perspectives and ways of valuing. The value creator must experiment with different perspectives and move back and forth between different ways of valuing to better understand the influences of external factors and, based on this, to be truly liberated from external constraints in order to create her own values and reach a state of autonomy.

Moreover, the pursuit of knowledge is motivated by a deeper, often unconscious drive for control, influence, and domination. This is evident in his work *Beyond Good and Evil*, where he discusses how the will to power manifests in our interpretations of the world. "Supposing that nothing else is 'given' as real but our world of desires and passions, that we cannot sink or rise to any other 'reality' but just that of our impulses—for thinking is only a relation of these impulses to one another" (*BGE* 21). Nietzsche here suggests that our entire perception and knowledge (our 'thinking') of the world are essentially interpretations shaped by our desires and passions, which are expressions of the will to power. To pursue power, knowing, as an act of interpreting, tries to subordinate nature to our will. Consequently, the pursuit of knowledge functions as a means of extending our power over nature, which is experienced as pleasure. As he writes, his aim is "to make knowledge the *most powerful* affect" (*SL*: 77). Knowing in this way contributes directly to an increase in power.

An example is *GM* II. Genealogy, as Nietzsche employs it, is a method that critically investigates the historical underpinnings and evolution of moral values, beliefs, and consequently, knowledge itself. This approach is aimed not simply at uncovering factual historical details but at understanding how these facts have been interpreted and the power dynamics that influenced these interpretations. Such an analysis is deeply rooted in the concept of the will to power, asserting that our pursuits of knowledge are expressions of our drive to assert control and extend our influence. By delving into the origins of values through genealogy, we aim to master these origins, thereby us having the ability to liberate ourselves from their dominance. This mastery provides a clearer lens through which to grasp the genesis and development of Christian morality, enabling us to critically evaluate and potentially discard its hold over modern culture.

Also, for Nietzsche, knowledge is not merely intellectual but existential. He explores the idea that gaining knowledge can lead to suffering, as it often involves disillusionment and the dismantling of comforting illusions. This is echoed in his parable of the madman in *The Gay Science*, who proclaims the death of God and the profound existential implications of this "knowledge" for modern humans. I will interpret this point in detail when talking about *amor fati*.

So far, I have articulated the role knowledge plays in value creation by helping humans get rid of external influences imposed on us. However, mere knowledge of external determination does not suffice for creating values. Rather, one also requires knowledge of oneself to discover one's inner principle expressive of one's nature. This idea connects directly to the origin of knowledge, which, through Nietzsche's concept of internalisation, expands the soul's depth, breadth, and height, creating a reflective inner realm. In other words, the nature of this soul—formed through internalisation—is not only conducive to acquiring knowledge of the external world but also cultivating reflective capacities. In this way, the origin of knowledge itself inherently facilitates the process of self-exploration, which enables individuals to dive into their own inner realities and discover the principles that truly express their nature. Moreover, in *GS* 335, when Nietzsche talks about our aim to become who we are and create ourselves, he writes that we should "*limit* ourselves to the purification of our opinions and valuations and to the *creation of our own new* tables of what is good" and to do so requires us to"become the best learners and discoverers of everything that is lawful and necessary in the world: we must become *physicists* in order to be able to be *creators* in this sense…" (*GS* 335). Nietzsche does not explicitly limit the "discoverers of everything that is lawful and necessary" to the external world. While the phrase in *GS* 335 emphasises understanding the natural laws and physical constraints that shape the world, it also leaves room for an application to the discovery of one's inner nature. Just as one must learn the lawful necessities of the world to navigate external reality, one must uncover the necessities within oneself to transcend self-ignorance and obtain the ability to create values that align with one's inner nature.

This understanding of knowledge about one's inner self aligns with Nietzsche's broader view of knowledge as interpretation and an expression of the will to power. Coming to know oneself is not simply a matter of uncovering static truths about one's nature but is an active process of self-interpretation. In this process, the individual does not passively observe or analyse objective facts about themselves as if these truths existed independently of their lived reality. Instead, they actively shape their understanding in a way that affirms their power and potential. Through this act of self-interpretation, one gains a deeper understanding of one's nature and develops the capacity to determine how values are created for oneself, because self-knowledge is not achieved through detached, "objective" observation, as though one could step outside one's own existence to examine it from an external perspective, a view that already presuppose a prior notion of freedom. Rather, it is grounded in the lived experience of one's existence. Ultimately, the process of external knowledge and self-discovery provides the condition for an independent and self-determined existence. Yet knowledge alone does not suffice for autonomy: it secures independence but does not in itself ground value creation. To reach Nietzsche's positive conception of free will, a further element is required—an ontological state toward life itself. This brings us to the second feature cultivated in the development of moral responsibility: *amor fati.*

**5.2 *Amor Fati***

In Chapter 4, I argued that noble value creation arises from an unconscious self-expression grounded in one's essential nature, which embodies what one is fated to be. Knowledge of oneself and the external world, as discussed in the previous sections, provides the condition for independence from external pressures and expectations. However, Nietzsche demands more than epistemological clarity: he requires an ontological state, a profound orientation toward one's own existence, because although knowledge secures a degree of independence, it cannot by itself enable one to affirm life positively or embrace necessity. For this reason, Nietzsche introduces *amor fati* as the ontological state that transforms mere independence into a creative and life-affirming orientation toward fate. This is vividly articulated in Nietzsche's *Twilight of the Idols*, where he writes that "One is necessary, one is a piece of fate, one belongs to the whole, one *is* in the whole…" (*TI* VI: 6). Since one's essence is fated to be as such, then to become autonomous, one needs to determine values for oneself according to what one is fated to be essentially, rather than external truths which one may discover. One, in this way, becomes a piece of fatefulness in pursuit of autonomy.

Given the close connection between autonomy and the acceptance of necessity, and the crucial role autonomy plays as one of the essential elements constituting free will, Nietzsche envisions a freedom that assents to fate as a condition of its own possibility. However, Nietzsche also maintains that the highest form of freedom for an individual does not reside merely in recognising one's inner power or "who he is," but in a complete identification with existence as a whole. This identification unfolds as an existential state toward life in its totality, which makes freedom transcend mere acceptance and moves towards a total embracing of all aspects of reality, including those that are challenging or unchangeable, in a manner that reflects an attunement to the inevitability of existence. In other words, freedom at its highest is realised in and through a profound affirmation and love of one's fate, an idea Nietzsche himself echoes in Goethe's sentiments in *Twilight of the Idols*: "A spirit thus emancipated stands in the midst of the universe with a joyful and trusting fatalism, in the faith that only what is separate and individual may be rejected, that in the totality everything is redeemed and affirmed—he no longer denies…." (*TI* IX: 49). As Nietzsche puts it, it is not merely the acceptance of fate, but the theme of *amor fati* (love of fate) that is his "formula for greatness in a human being" (*EH* 258), which is "the highest state a philosopher can attain" (*WP* 1041). Crucially, *amor fati* is not an instance of willing to love or a deliberate choice of emotional attachment; rather, it is an existential attitude, a state in which one affirms the givenness of one's own life and the inevitability of one's essential nature. With knowledge and the affirmation of necessity in place, the individual is able to determine her own principles for action, act in accordance with her essential nature, and integrates her understanding of fate into her self-governance. Nietzsche's ideal figures, such as the philosopher of the future and the sovereign individual, whom he thinks to bear great and noble morality, embody the principle of *amor fati* and integrate it into their approach to the re-creation of values and the development of a life-affirming philosophy.

Here one might ask: how does one come to love their fate? To be more specific, how can one love something that necessarily contains negative aspects? As Nietzsche puts it, "one will see that the problem is that of the meaning of suffering" (*WP* 1052). In several places, he emphasises the possibility that our fate is to suffer endlessly. As a matter of fact, fate inevitably entails suffering and unhappiness for every one of us. We are all bound to farewell, to witnessing loved ones dying and leaving us behind. But this is not to say that fate consists only of negativity. There certainly exist some positive aspects that illuminate life, but similar to what Schopenhauer claims to be, Nietzsche holds that the underlying substance of one's life is marked by negativity. Even if one's life could be devoid of any unhappiness, one will face ageing and death from the moment of birth. Therefore, there seems to be an explicit tension between the nature of love and the negative value of its object. To love fate, one would have to accept the negative possibilities intrinsic to this lived experience, which is profoundly difficult given our human nature.

Moreover, the fate we are called to love is not only one side, but the whole in its entirety. That is to say, loving of fate is deeply connected with eternal recurrence. We should "perceive not merely the necessity of those sides of existence hitherto denied, but their desirability; not their desirability merely in relation to the sides hitherto affirmed, but *for their own sake*" (*WP* 1041). As Nietzsche clearly shows when talking about human greatness:

> My formula for greatness in a human being is *amor fati*: that one wants nothing to be different, not forward, not backward, not in all eternity. Not merely bear what is necessary, still less conceal it—all idealism is mendacity in the face of what is necessary—but *love it*. (*EH* 258)

The attitude towards fate, for Nietzsche, is not merely "bearing it", which implies a passive adaptation to the environment, nor is it to "conceal it" by minimising the reality of suffering. Rather, it is to embrace fate from a synchronic standpoint, which contains necessity in relation to every aspect of life, from present, to forward, and to backward. Moreover, Nietzsche further claim that *amor fati*, which he calls "the highest state a philosopher can attain", is "to stand in a Dionysian relationship to existence", which "wants the eternal circulation—the same things, the same logic and illogic of entanglements" (*WP* 1041). Therefore, *amor fati* is not merely about loving what is necessary and fated to be, nor is it simply an attitude toward the future; it is an orientation toward existence as a whole, an affirmation of life in its recurrence.

To address how to love fate, Beatrice Han-Pile (2011) proposes a distinction between two forms of love: erotic love and agapic love. Han-Pile explains that the difference between the two corresponds broadly to Christian and Greek forms of love. Erotic love arises from the perceived value of its objects: one loves because one sees something desirable in the beloved. Just like a luxury item whose appeal stems less from its intrinsic qualities than from the prestige of the brand it carries, our love for it arises from the value it represents rather than from its inherent features. By contrast, agapic love is not grounded in the prior worth of its object; instead, it is through the experience of loving that its significance becomes manifest. A simple, even ordinary, object—say, a pen, can become precious if given by a loved one, not because of its intrinsic properties, but because love itself confers significance. In this sense, it is not that we love something because we value it, but we value something because we love it.

Now consider how erotic love, which emphasises the value of its objects, can face the continual return of fate. Many Nietzsche scholars (Nehamas, 1985; Clark, 1990; Register, 2006, etc.) interpret eternal recurrence as a thought experiment to test our ability to love fate from a certain standpoint and help to develop a positive perspective on what is necessary. Nehamas discusses how eternal recurrence serves as a thought experiment to test our ability to fully affirm life, including its challenging aspects (Nehamas, 1985: 153, 156, 194). Similarly, Register suggests that eternal recurrence is a test of one's capacity to embrace life's necessities and challenges, thus fostering a positive perspective on fate (Reginster, 2006: 90-92). Clark understands the demon's challenge similar to the question: "if you had to do it all over again, would you marry me again?" (Clark, 1990: 269). Underlying these accounts is the idea that eternal recurrence lays out a situation about a demon telling us that we have to live our life again and again, presenting a standpoint detached from our lived experience that we are asked to evaluate our life as a whole, and then tries to get the conclusion from the experiment, which represents our attitudes towards life. From this standpoint, if we conclude that our life is worth living again and again, our love of fate will be proven. However, there is a serious problem in this interpretation concerning the essence of love and value: value is not identical with love. The fact that our life is worth living again and again does not necessarily give the conclusion that we should love it. As Han-Pile cites Kant's saying, "*love* is a matter of *feeling*, not of willing, and I cannot love because I *will* to, still less because I *ought* to" (Kant, 1996: 530, cited in Han-Pile 2011: 231). Indeed, there is a gap between loving and valuation, because love implies a valuation of its object, but cannot be reduced to its valuation. If something or someone can evoke my love for it, I will certainly value it, but I can value it without loving it. For example, pupils may value going to school because they know that studying is important, but they can still hate it for taking away their playtime. Similarly, when confronted with the demon's question, I can value my life positively and want it to return over and over, but this conclusion may be produced by rational consideration, not out of love. Erotic love, since it arises from the perceived worth of its object, cannot capture the kind of love at stake in amor fati. For Nietzsche, the love of fate is not grounded in a consequentialist appraisal of life's value, but consists in an affective state in which life is embraced as such. We love our fate not because of its perceived worth from a consequentialist standpoint. We love it just because we love it.

What remains now is the agapic love for fate. Nietzsche himself recognises the gap between love and worth, claiming that "it is impossible to teach love", which lends support to the idea that the love of fate he envisions is an agapic one. For him, "it is love alone that can bestow on the soul, not only a clear, discriminating and self-contemptuous view of itself, but also the desire to look beyond itself and to seek with all its might for a higher self as yet still concealed from it" (*UM* 162-3). In this picture, love does not follow from the object's value; rather, it is love that brings value into view, disclosing possibilities of transformation and elevation. We love our fate not because of its inherent attractiveness or worth. On the contrary, even with all its suffering and unhappiness, our love per se motivates us to overcome life's negative aspects, and through such love, the world is transfigured into something beautiful. With love, "I shall be one of those who make things beautiful. *Amor fati*: let that be my love henceforth!" (*GS* 276).

Now the problem is, how does this agapic love come into existence? Han-Pile claims that, for Nietzsche, the fact that loving a seemingly negative fate might be desirable does not provide an argument capable of persuading us into such a state of love. Rather, it is a lived experience, which has to be undergone to be fully understood. To capture why agapic love cannot be rationally willed into existence, Han-Pile ingeniously introduces the conception of *mediopassive modality* (Han-Pile, 2011: 233). Different from most Indo-European languages, which distinguish only active and passive modes, ancient Greek had a third mode, in which the agent participates in an action without fully controlling it. In other words, one is both active and passive in this action. In this sense, the peculiar modality of the expression "let that be my love henceforth" can be read as an action, like letting it be, which is neither fully active nor passive, indicating that one participates in the process of love without being in full control of this love. Seen in this light, *amor fati* differs from the wilful erotic pursuit of love, which assumes that the active appreciation of the objects' value can generate love. The agapic lover does not approach fate from a detached standpoint and actively look for reasons to make fate lovable. Rather, she finds herself in a state where love already transforms how life appears, disclosing its value from within. The transfiguration of life's negative aspects is thus not an argument for why one ought to love fate, but observations and descriptions of what it is like to be in such a agapic state—one that is lived through rather than abstractly reasoned about. Importantly, this state is neither a matter of random luck nor of deliberate choice: not everyone can finally achieve it, but it becomes possible for those who cultivate a deep understanding of themselves and the world. In this sense, *amor fati* arises not because one decides to love, but because such understanding allows this love to emerge and permeate one's state toward existence. Therefore, *amor fati* cannot be wilfully motivated and come into existence by reasonably changing one's perspective; it is a modality of existence that one inhabits.

Moreover, if *amor fati* is not willed into being but must be lived through, then, what kind of lived experiences can open the way to such love? Nietzsche identifies suffering as one such experience, because in response to suffering, certain features may emerge that transfigure this negative experience into a beneficial thing. An essential feature among them is courage, which is not a classical virtue cultivated through habituation, but rather the existential capacity to resist fear and remain open to life's challenges that can be developed in the face of suffering. Human beings, driven by the natural instinct to seek benefit and avoid harm, often respond to suffering by shielding themselves, searching for escape routes, or developing resentment towards their current circumstances, feeling that they are treated in an undeserved manner. When confronted with suffering, people might choose to shield themselves and retreat from the real world into their inner world (as the Stoics might), or, like Nietzsche's Christians, harbour resentment towards the masters of reality, abandoning the real world for a fantasy in which they hope to be healed (*GM* II). This situation changes when courage is developed to counteract the rise of fear brought about by suffering. As Nietzsche remarks, *amor fati* is not a passive waiting, but an active striving "against the pessimism of weariness with life" (*HH* II: 212-3). This striving should not be understood as a deliberate, willed choice; rather, it manifests as the expression of strength in the face of obstacles, arising naturally through lived experience. Courage, as an expression of strength, can enable individuals to confront the pain of suffering and to draw meaning from it. Equipped with courage, individuals are less likely to succumb to self-denial, trying to run away from difficulties and tend to pursue other alternatives, because they come to realise that what is truly valuable in life requires us to open ourselves to all possibilities. Refusing all contact with the outside world may seem to protect oneself, but in reality, it cuts off the possibility of deriving value from life. Indeed, suffering brings sadness and discomfort, but through courage one opens oneself to suffering as part of existence, and thereby also to the beauty of life. As Nietzsche says, "if you refuse to let your own suffering lie upon you for an hour and if you constantly try to prevent and forestall all possible stress way ahead of time; if you experience suffering and displeasure as evil, hateful, worthy of annihilation, and as a defect of existence, then it is clear that besides your religion of pity you also harbour another religion in your heart that is perhaps the mother of the religion of pity: the religion of comfortableness. How little you know of human happiness, you comfortable and benevolent people, for happiness and unhappiness are sisters and even twins that either grow up together or, as in your case, remain small together" (*GS* 269-270). A love for fate can only be fully realised through lived experience, and it is courage—this existential openness to suffering and risk—that makes such an experience possible.

From this, it becomes evident that *amor fati* does not emerge in isolation; it evolves alongside the development of moral responsibility, whereby individuals come to experience their own existence through moral practice. For Nietzsche, moral responsibility develops as individuals move beyond externally imposed notions of guilt—such as the Christian conception of guilt, which Nietzsche criticises for burdening the individual with inescapable debt—towards a self-determined form of accountability. Although traditional notions of moral responsibility inevitably involve suffering and negative emotions like guilt, these "negative" experiences become integral to the lived experience and provide the ground for a revaluation of life. To love every aspect of life, then, is not merely to endure mistakes and misfortunes but to affirm them as constitutive of one's existence. Through the ongoing practice of affirmation, the grip of denial and negative emotions gradually loses its hold. In this process, *amor fati* functions as a philosophical practice that transforms suffering into a source of strength: the strong will is shown not in erasing the past but in affirming it as the condition for creative becoming. At the heart of this creative attitude lies *amor fati*: it demands that individuals accept the inevitability of life, including their limitations and failures, and transform these limitations into possibilities for new value-creation. Through this experience, individuals draw strength from past suffering, open themselves to discomforts, assume greater responsibility for the future, and thereby shape new values in an active, life-affirming way.

In summary, Nietzsche's conception of free will is not merely to accept what one is essentially fated to be, but to affirm fate in its entirety. Free will, for Nietzsche, consists in the lived capacity to love fate, that is, *amor fati*! It is acknowledged that fate can be negative and perhaps the very essence of fate is suffering, but the loving of fate, as Han-Pile emphasises, is agapic in nature: it does not arise from the prior value of what is loved, but from a state that affirms life unconditionally. We do not love fate because of its perceived values that attract us and makes us feel it is worthy of love. We love fate because we are in a state to understand the lovable aspects of fate arising not from any external incentive or imposed reason, but from the individual's lived orientation toward life itself. In loving fate, we transfigure its value, bestowing worth upon the whole of our lived experience. Such transfiguration is possible through the experience of suffering, which helps to cultivate certain qualities that give rise to dispositions that open the possibility of overcoming negative valuation even if fate were perceived as a fully negative object. These qualities, like courage and calmness, do not protect us from pains in life, nor diminish the painful experience, but enable us to directly embrace these experiences as fated, and transform ourselves through this lived experience, thereby gaining more possibilities and value from our encounter with suffering. Thus, *amor fati* is not an act of rational choice but an ontological state that manifests the strength of will: the ability to transform even what appears hostile or negative into a source of affirmation. In this sense, to love one's fate is precisely what distinguishes a strong will—not because one has chosen it over alternatives, but because one has the integrative power to affirm necessity as the very medium of value-creation. Not rejecting life but welcoming it in its totality—this is the true meaning of *amor fati*.

### 5.3 Autonomy

Having clarified *amor fati* as the ontological state that grounds value creation, we can now turn to autonomy. For Nietzsche, autonomy is not merely independence from external constraints but the positive condition that arises once independence and *amor fati* converge, enabling the individual to establish their own principles of action. Nietzsche's conception of autonomy is central to his idea of free will, but the meaning of autonomy is not yet clear. Nietzsche's use of the term "autonomy" or "autonomous" is indeed rare, reflecting his unique approach to the themes commonly associated with these terms. He typically discusses concepts related to self-governance, freedom, and independence using other terms like "sovereignty" or "self-overcoming." However, after a review, it seems that direct references to "autonomy" or "autonomous" are not prevalent in the most cited translations of his works (in German "autonom") . Yet, in the second essay of *On the Genealogy of Morals*, Nietzsche directly refers to this term and ascribes this concept to the sovereign individual:

> If we place ourselves at the end of this tremendous process, where the tree at last brings forth fruit, where society and the morality of custom at last reveal *what* they have simply been the means to: then we discover that the ripest fruit is the *sovereign individual*, like only to himself, liberated again from morality of custom, autonomous and supramoral (for 'autonomous' and 'moral' are mutually exclusive), in short, the man who has his own independent, protracted will and the *right to make promises*—and in him a proud consciousness, quivering in every muscle, of *what* has at length been achieved and become flesh in him, a consciousness of his own power and freedom, a sensation of mankind come to completion. This emancipated individual, with the actual *right* to make promises, this master of a *free* will, this sovereign man—how should he not be aware…of how this mastery over himself also necessarily gives him mastery over circumstances, over nature, and over all more short-willed and unreliable creatures? (*GM* II: 2)

It is clear that in this passage, Nietzsche links autonomy to the liberation from traditional morality, independence, free will and a consciousness of one's power and freedom. Since he deliberately emphasises that "'autonomous' and 'moral' are mutually exclusive", it can be inferred from this saying that autonomy requires the liberation from external constraints (such as Christian morality) and presupposes independence as a necessary condition. However, being free from external constraints is not sufficient for autonomy, because without setting up principles for one's actions, how one determines value and acts on this basis will be arbitrary, and naturally, one can talk even less about autonomy. This shows that autonomy for Nietzsche must be more than mere liberation; it requires a new evaluative stance rooted in one's own nature. As is shown as he continues in the same passage:

> The "free" man, the possessor of a protracted and unbreakable will, also possesses his *measure of value:* looking out upon others from himself, he honors or he despises; and just as he is bound to honor his peers, the strong and reliable (those with the *right* to make promises)…The proud awareness of the extraordinary privilege of *responsibility,* the consciousness of this rare freedom, this power over oneself and over fate, has in his case penetrated to the profoundest depths and become instinct, the dominating instinct. (*GM* II: 2)

As he puts it, the free person who possesses autonomy has his own "*measure of value*" and makes it his dominating instinct, leading to his constant adherence to his own measure of value ("unbreakable will" and "the right to make promises"). Knowledge of both external and internal obstacles contributes to this process by helping individuals better understand the rules and complexity of the surrounding world and their neglect of their own inner nature. Yet this measure of value is not merely defensive or independent; it is essentially creative. The autonomous individual does not only distance herself from external chaos but also reinterprets it through her own measure of value, continuously re-evaluating and re-signifying events. Here *amor fati*, which is described as "the power over fate", becomes central: to love one's fate is to exercise sovereignty over necessity itself and incorporate even suffering and misfortune into a positive framework of meaning. Such an existential state enables the free spirit to integrate the negative into a creative process of valuation, thereby forming an authoritative principle of action grounded in the dynamic unfolding of her own nature—understood as the capacity for value-creation, re-interpretation, and the affirmation of life. In this way, autonomy is not simply independence from external influence but the capacity to generate a life-affirming horizon of value through the creative embrace of necessity.

Although Nietzsche rarely uses the terms "autonomy" (Autonomie) and "autonomous" (autonom) , his philosophy often engages with the theme indirectly. Many of his reflections on independence, solitude, and the struggle against external norms can be read as alternate ways of expressing autonomy. For example, he praises those far above the average as possessing "high and independent spirituality" and "the will to stand alone" (*BGE* 201), describes "free spirits" as "jealous friends of *solitude*" (*BGE* 44), and even identifies greatness with "wanting to be by oneself, being able to be different, standing alone and having to live independently" (*BGE* 212). These passages show that Nietzsche often frames autonomy in terms of independence. For Nietzsche, attaining autonomy and becoming noble is a struggle, a war against others' views, options, and expectations to free oneself from such external influences. Yet independence here functions only as a necessary groundwork: it frees the individual from the weight of external norms and expectations, but it does not yet suffice for autonomy in the full sense.

What, then, gives autonomy its positive content? For Nietzsche, what is required in addition is the establishment of an inner *measure of value*, a principle by which the individual reinterprets and re-evaluates existence. This is why Nietzsche does not advocate a simple return to the Greeks or the Stoics, however admirable their virtues. The Greeks developed a relatively "healthy" form of responsibility, but they did not possess the capacity for self-legislation of values; the Stoics, likewise, treated external order as the arbiter of meaning. By contrast, Nietzsche anticipates future philosophers who, beyond merely rejecting external judgements, create new values. Such creation is not a momentary decision but an existential state, epitomised in *amor fati*: the power to love one's fate, to integrate even suffering and misfortune into a positive horizon of meaning. In this sense, autonomy draws its substance from the constitutive aim of valuing (CA-V), what we discussed before: to generate new perspectives, to transform what is given, and to intensify life. Autonomy is achieved not simply by standing apart, but by creatively affirming necessity as the material through which one's evaluative standpoint is forged. This explains why in *GM II:2* Nietzsche explicitly links autonomy not just to liberation from custom but to the possession of a "protracted and unbreakable will" and the right to make promises, thereby grounding responsibility in an inner principle of valuation. Autonomy thus emerges at the point where independence and value-creation converge. This convergence explains how Nietzsche’s sovereign individual can be intelligible as a bearer of responsibility: not because she has escaped all conditioning, but because she embodies a stable evaluative standpoint through which she can be held answerable for her actions.

This brings us to the crucial question: if values are self-created in this way, where does their authority come from? Or, to put it another way, what distinguishes autonomous valuation from arbitrary preference? As discussed in Chapter 4, the debate between value objectivism and subjectivism offers two unsatisfactory poles. Value objectivism grounds authority in correspondence to objective truths about value. One is normatively bound to this principle; her actions are justified if they are consistent with objective truth. By contrast, subjectivism reduces authority to individual endorsement. As I argued, Nietzsche's account points toward a middle ground between objectivism and subjectivism. The authority of self-created principles lies in their constitutive role, which is intrinsic to one's power and personal values in the sense that one cannot fail to obey those principles because of the person one is. In other words, the normative force of these principles is internal and constitutive: it arises from the very structure of the valuing activity and the identity of the individual, not from any external truth or arbitrary imposition. An individual's failure to follow those principles comes at the expense of her identity, and her successful observation of these lawful forces is not merely consistent behaviour, but an expression of her strength, her capacity to affirm and extend life through creative valuation. In this sense, autonomy is grounded not in external truth or arbitrary choice, but in the generative activity of valuing itself—an activity whose highest expression is *amor fati*.

The noble type of human is such an example of an autonomous agent who determines values and principles for himself out of his own inner strength. As Nietzsche writes, "The noble type of man experiences *itself* as determining values; it does not need approval; it judges, 'what is harmful to me is harmful in itself'; it knows itself to be that which first accords honor to things; it is *value-creating"* (*BGE* 260). A similar picture applies to the philosopher of the future, who "demands of himself a judgement, a Yes or No, not about the sciences but about life and the value of life" (*BGE* 205), and the sovereign individual, who "also possesses his *measure of value*" (*GM* II: 2). Unlike the Stoics, they do not look outward to a supposed inherent order of the world. Rather, they locate the normativity in self-knowledge and self-legislation, and thus set their principles in accordance with their own power and future development. In this sense, an autonomous individual does not rely on the external world to validate her values, nor does she attempt to impose her value judgements as universally binding for all. Instead, her standpoint is internally grounded: she maintains her independence and establishes her own principles for action that affirm her own nature and creatively transform her circumstances. The resulting values and normative principles are not "objective" in the sense of being universally shared, but neither are they arbitrary. They carry binding force because they embody a stable evaluative standpoint—one that integrates affirmation, transformation, and continuity of self. This form of stability does not imply rigidity or immutability; rather, it signifies a durable core of evaluative identity that allows the individual to creatively transform her circumstances while remaining faithful to her fundamental self. In this way, the autonomous agent retains authority and is capable of holding herself answerable for her commitments, even amidst the ongoing processes of reflection, experimentation, and self-reinvention.

To sum up, autonomy in Nietzsche's sense arises only once knowledge and *amor fati* are integrated into a unified statge. Knowledge of external constraints enables the individual to maintain independence from inherited norms and resist domination by the crowd. *Amor fati* gives this independence its positive content: it is not a rational posture one might simply adopt, but an ontological lived state in which necessity itself is embraced as the ground of affirmation. On this basis, value creation becomes possible: the individual does not merely negate existing norms but reinterprets and transforms what is given into principles expressive of her power. Autonomy therefore marks the transition from separation to genuine self-legislation. It is not reducible to solitude or independence, nor exhausted by creativity alone, but signifies the point at which independence and value-creation converge into a stable evaluative standpoint. In this sense, the autonomous individual exemplifies Nietzsche's reworking of the Stoic schema: while the Stoics traced their actions back to external order, Nietzsche roots his understanding of fate in a lived state toward necessity itself. Autonomy thus designates the culmination of this development—an existential condition in which the individual becomes genuinely self-governing, and thereby a bearer of strong will.

## 6. Conclusion

In this chapter, I first clarified the relationship between free will and moral responsibility. Free will and moral responsibility are often conflated, but actually they cannot be assumed to be used interchangeably. The significance of this chapter, therefore, lies in further outlining Nietzsche's idea of free will based on his moral perspective. As I argued, free will is not the same as freedom, but is the combination of freedom and will, bringing about the will in a strong sense that can affirm itself and resist external influences. Then, to clarify the conception of freedom and will, I got inspiration from the ancient notions of free will, particularly the Stoics' attempt to address the apparent tension between free will and fate. Although Nietzsche explicitly opposed the philosophy of the Stoics, their articulation of free will as a competence provides us with a useful contrast. For the Stoics , the will is an ability to choose to assent to appropriate impressions and act according to this willing, whereas freedom entails the refusal of external forces that would unwillingly affect our choice-making. Wise persons are those taking care of their self-maintenance ability based on their understanding of the order of the world and who choose to assent to appropriate impressions that contribute to the overall harmony of universe without coercion or hindrance. In this regard, the moment of liberation occurs just when one understands without regret that nothing could have been different as it is, being a part of the whole necessity. Nietzsche is against the Stoics' return to an inner state, detached from the activation of emotions and the experience of sufferings. He further develops the Stoics' idea and claims that liberation is not detachment from the passions but their affirmation, not resignation but active engagement with necessity. In this way, Nietzsche lays the ground for a conception of free will as the highest expression of life. Nietzsche's free will is thus best understood as a will full of strength that gives an individual the ability to create values and assume responsibility for herself in one's lived experience, rather than a conceptual notion of free will in the traditional sense. Its strength does not consist in metaphysical control or alternative possibilities, but in the existential integration of knowledge, affirmation, and value creation. **Knowledge** enables one to maintain independence from external constraints; ***amor fati***, understood as a lived ontological state—secures the affirmation of life and provides the ground for value creation; and together these conditions make possible **autonomy**, the mastery of one's life through the creation of values expressive of one's power. Such an autonomous individual willingly and proudly assumes responsibility, not because she could have acted otherwise, but because she actively affirms her being and her deeds as necessary. Based on the autonomy she attains, one not only assents to but also positively affirms the necessity of fate, which motivates her to bear full responsibility for what she has done and be proud of her being as such. Even if the full scope of Nietzsche's view of freedom remains contested, the account I have given identifies several key components and shows why Nietzsche reconfigures free will as strong will: a will that integrates knowledge, affirmation, and creation into a form of autonomy that allows the individual to embrace life and bear responsibility in the most affirmative sense.

# Bibliography

## 1. Nietzsche's Texts

AC *Anti-Christ*, Trans. Trans. R. J. Hollingdale, London: Penguin Books, 1990

BGE *Beyond Good and Evil*, Trans. Walter Kaufmann, New York: Vintage, 1996.

D *Daybreak: Thoughts on the Prejudices of Morality*, Trans. R. J. Hollingdale, Cambridge: Cambridge University Press, 1997.

EH *Ecce Homo*, Trans. Walter Kaufmann and R. J. Hollingdale, New York: Vintage, 1989.

GS *The Gay Science*, Trans. Walter Kaufmann, New York: Vintage, 1974.

GM *On the Genealogy of Morals*, Trans. Walter Kaufmann and R. J. Hollingdale, New York: Vintage, 1989.

HH *Human, All Too Human*, Trans. R. J. Hollingdale, Cambridge: Cambridge University Press, 1986.

LN *Writings from the Late Notebooks*, ed. Rüdiger Bittner. Cambridge: Cambridge University Press, 2003.

SL *Selected Letters of Friedrich Nietzsche*, trans. Christopher Middleton. Indianapolis: Hackett, 1996.

UM *Untimely Meditations*, R. J. Hollingdale, Cambridge: Cambridge University Press, 1997.

TI *Twilight of the Idols*, Trans. Trans. R. J. Hollingdale, London: Penguin Books, 1990.

WP *The Will to Power*, Trans. Walter Kaufmann and R. J. Hollingdale, New York: Vintage, 1968.

Z *Thus Spoke Zarathustra*, Trans. Walter Kaufmann, London: Penguin Books, 1978.

eKGWB *Digitale Kritische Gesamtausgabe Werke und Briefe*, ed. Paolo D'Iorio, http://nietzschesource.org, which is based on KGW

**2. Others Works**

Acampora, C. D. (2006). On sovereignty and overhumanity: Why it matters how we read Nietzsche's Genealogy II: 2. *International Studies in Philosophy*, *36*(3), 127-145.

Anderson, R. L. (2022). Nietzschean autonomy and the meaning of the "sovereign individual". *Philosophy and Phenomenological Research, 105*, 362–384. https://doi.org/10.1111/phpr.12824

Arpaly, N. (2006). *Merit, Meaning and Human Bondage: An Essay on Free Will*, New Jersey: Princeton University Press.

Baumeister, R., Masicampo, E.j., & Vohs, K. (2011). Do conscious thoughts cause behavior? *Annual review of Psychology*, 331-361. https://doi.org/10.1037/14341-007

Bengtson, A. B. (2020). The concept of moral responsibility and some very general facts of nature: A Wittgensteinian approach to Strawson's "Freedom and Resentment" (Doctoral dissertation, KU Leuven).

Boehm, C. (2012). *Moral Origins.* New York: Basic Books.

Bourdieu, P & Wacquant, L. (1992). *An Invitation to Reflexive Sociology.* Polity Press.

Chalmers, D. J. (2011). Verbal disputes. *The Philosophical Review.* Durham: Duke University Press. *120* (4): 515–566. https://doi.org/10.1215/00318108-1334478

Clark, M. (1990). *Nietzsche on Truth and Philosophy*. Cambridge: Cambridge University Press.

Clark, M & Dudrick, D. (2009). Nietzsche on the will: An analysis of *BGE* 19. *Nietzsche on Freedom and Autonomy*, ed. K. Gemes & S. May. Oxford: Oxford University Press.

Cushman, F., Young, L., & Greene, J. D. (2010). Multi-system moral psychology. *The moral psychology handbook*, ed. J. M. Doris & Moral Psychology Research Group. Oxford: Oxford University Press.

de WAAL, F. B. M. (1991). The Chimpanzee's sense of social regularity and its relation to the human sense of justice. *American Behavioral Scientist*, *34*(3), 335-349. https://doi.org/10.1177/0002764291034003005

(2006). *Primates and philosophers: How morality evolved.* Princeton University Press.

Drapela, N. (2019). A matter of taste: Nietzsche and the structure of affective response. *Inquiry*, *63*(1), 85–103. https://doi.org/10.1080/0020174X.2019.1669979

Figdor, C., & Phelan, M. (2015). Is free will necessary for moral responsibility?: A case for rethinking their relationship and the design of experimental studies in moral psychology. *Mind & Language, 30*(5), 603–627. https://doi.org/10.1111/mila.12092

Fischer, J. M. (2006). *My Way: Essays on Moral Responsibility.* New York: Oxford University Press.

Fischer, J. M., & Ravizza, M. (Eds.). (1993). *Perspectives on moral responsibility*. Cornell University Press.

Foot, P. (1972). Morality as a system of hypothetical imperatives. *The Philosophical Review, 81*(3), 305–316. https://doi.org/10.2307/2184328

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*(1), 5–20. https://doi.org/10.2307/2024717

Frede, M. (1987). *Essays in Ancient Philosophy* (NED-New edition). University of Minnesota Press. http://www.jstor.org/stable/10.5749/j.cttttpfd

(2011). *A Free Will: Origins of the Notion in Ancient Thought*. London: University of California Press.

Gemes, K. (2009). Nietzsche on free will, autonomy and the sovereign individual. *Nietzsche on Freedom and Autonomy*, ed. K. Gemes & S. May. Oxford: Oxford University Press.

Gettier, Edmund L. (1963). Is Justified True Belief Knowledge? *Analysis, 23*(6):121-123.

Greenspan, P. S. (1992). Subjective guilt and responsibility. *Mind*, *101*(402), 287–303. http://www.jstor.org/stable/2254336

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judGMent. *Psychological Review, 108*(4), 814–834. https://doi.org/ 10.1037/0033-295X.108.4.814

Han-Pile, B. (2011). Nietzsche and Amor Fati. *European Journal of Philosophy*, 19, 224-261. https://doi.org/10.1111/j.1468-0378.2009.00380.x

(2019). 'The doing is everything': a middle-voiced reading of agency in Nietzsche. *Inquiry*, *63*(1), 42–64. https://doi.org/10.1080/0020174X.2019.1669977

Harper, A. (2012). Nietzsche on value creation (Doctoral Dissertation, University of Illinois at Urbana-Champaign).

Hatab, L. J. (2008). *Nietzsche's "On the Genealogy of Morality": An Introduction.* Cambridge: Cambridge University Press.

Hoefer, C. (1996). The metaphysics of spacetime substantivalism. *The Journal of Philosophy*, *93*(1), 5–27.

Hughes, F. (2002). Nietzsche's janus perceptions and the construction of values. *Journal of the British Society for Phenomenology, 33*(2), 116-137.

Hussain, Nadeem J. Z. (2007). Honest illusion: Valuing for Nietzsche's free spirits. *Nietzsche and Morality*, eds. B. Leiter and N. Sinhababu. Oxford: Oxford University Press.

Janaway, C. (2006). Nietzsche on free will, autonomy and the sovereign individual. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *80*, 321–357. http://www.jstor.org/stable/4107046

(2007). *Beyond Selflessness*, Oxford: Oxford University Press.

(2009). Autonomy, affect, and the self in Nietzsche's project of genealogy. *Nietzsche on Freedom and Autonomy*, ed. K. Gemes & S. May. Oxford: Oxford University Press.

Kant, I. (1996). *The Metaphysics of Morals*, trans. M. J. Gregor, in *Immanuel Kant, Practical Philosophy*, trans. And ed. M. J. Gregor. Cambridge: Cambridge University Press.

Katsafanas, P. (2011). Deriving Ethics from Action: A Nietzschean Version of Constitutivism. *Philosophy and Phenomenological Research*, *83*(3), 620–660. http://www.jstor.org/stable/23210044

(2013). *Agency and the foundations of ethics: Nietzschean constitutivism*. Oxford University Press.

(2016). *The Nietzschean Self: Moral Psychology, Agency, and the Unconscious.* Oxford: Oxford University Press.

Katz, P. A., & Kofkin, J. A. (1997). Race, gender, and young children. *Developmental Psychopathology: Perspectives on Adjustment, Risk, and Disorder*, ed. S. S. Luthar, J. A. Burack, D. Cichetti, & J. Weisz. NY: Cambridge University Press.

Kornhuber, H. H., and Deecke, L. (1990). Readiness for movement—the Bereitschafts potential-story. *Curr. Contents Life Sci, 33*(14).

Leiter, B. (2002). *Routledge Philosophy Guidebook to Nietzsche on Morality.* London: Routledge.

(2007). Nietzsche's theory of the will. *Philosophers' Imprint, 7*, 1-15.

(2010). "Who is the 'sovereign individual'? Nietzsche on freedom," in *Cambridge Critical Guides to Nietzsche's On the Genealogy of Morality*, ed. S. May, Cambridge: Cambridge University Press.

(2011). Who is the "sovereign individual"? Nietzsche on freedom. *Cambridge Critical Guide to Nietzsche's On the Genealogy of Morality*, ed. Simon May. Cambridge: Cambridge University Press.

(2013). Moralities are a sign-language of the affects. *Social Philosophy and Policy*, *30*(1-2), 237-258. doi:10.1017/S0265052513000113

(2013). Moral skepticism and moral disagreement in Nietzsche. *Oxford Studies in Metaethics, 9*. University of Chicago, Public Law Working Paper No. 257. https://doi.org/10.2139/ssrn.1315061

(2015). *Nietzsche on Morality.* New York: Routledge.

(2019). *Moral Psychology with Nietzsche.* Oxford: Oxford University Press.

(2021). Disagreement, anti-realism about reasons, and inference to the best explanation. *Ethic Theory Moral Practice*.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will involuntary action. *Behav. Brain Sci,* 8(4), 529-539. doi:10.1017/S0140525X00044903

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly*, *2*(4), 245–264.

May, S. (2009). Nihilism and the free self. *Nietzsche on Freedom and Autonomy*, ed. K. Gemes & S. May. Oxford: Oxford University Press.

McKenna, M. (2009). Compatibilism. *The Stanford Encyclopedia of Philosophy,* Winter 2009 Edition, E. N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2009/entries/compatibilism/>.

(2012). *Conversation and Responsibility.* Oxford: Oxford University Press.

Mele, Alfred R. (2006). *Free will and Luck.* Oxford: Oxford University Press.

Mitchell, J. (2017). Nietzsche on Taste: Epistemic Privilege and Anti-Realism. *Inquiry*. https://doi.org/10.1080/0020174X.2016.1251166

Nahmias, E. (2010). Scientific challenges to free will. *A Companion to the Philosophy of Action*, ed. T. O'Connor, & C. Sandis, Wiley-Blackwell.

Nehamas, A. (1985). How one becomes what one is. *Nietzsche: Life as Literature.* Cambridge, Mass.: Harvard University Press.

Owen, D. (2009). Autonomy, self-respect, and self-love: Nietzsche on ethical agency. *Nietzsche on Freedom and Autonomy*, ed. K. Gemes & S. May. Oxford: Oxford University Press.

Pereboom, D. (2001). *Living Without Free Will.* Cambridge: Cambridge University Press.

(2016). *Free Will, Agency, and Meaning in Life.* Oxford: Oxford University Press.

Pigliucci, M. *How to Be a Stoic: Using Ancient Philosophy to Live a Modern Life*. Basic Books, 2017.

Pippin, R. (2009). How to overcome oneself: Nietzsche on freedom. *Nietzsche on Freedom and Autonomy*, ed. K. Gemes & S. May. Oxford: Oxford University Press.

Poellner, P. (2007). Affect, value and objectivity. *Nietzsche and Morality*. eds. B. Leiter and N. Sinhababu. Oxford: Oxford University Press.

(2009). Nietzschean freedom. *Nietzsche on Freedom and Autonomy*, ed. K. Gemes & S. May. Oxford: Oxford University Press.

Reginster, B. (2006). *The Affirmation of Life.* Cambridge: Harvard University Press.

(2017). What is the structure of *Genealogy of Morality* II? *Inquiry*, *61*(1), 1–20. https://doi.org/10.1080/0020174X.2017.1371833

Richardson, J. (2004). *Nietzsche's New Darwinism.* Oxford: Oxford University Press.

(2009). Nietzsche's freedom. *Nietzsche on Freedom and Autonomy*, ed. K. Gemes & S. May. Oxford: Oxford University Press.

(2020). *Nietzsche's Values.* Oxford: Oxford University Press.

Ridley, A. (2009). Nihilism and the free self. *Nietzsche on Freedom and Autonomy*, ed. K. Gemes & S. May. Oxford: Oxford University Press.

Risse, M. (2003). Origins of ressentiment and sources of normativity. *Nietzsche-Studien, 32*(1), 142–170. https://doi.org/10.1515/9783110179200.142

(2009). The eternal recurrence: A Freudian look at what Nietzsche took to be his greatest insight. *Nietzsche on Freedom and Autonomy*, ed. K. Gemes & S. May. Oxford: Oxford University Press.

Riccardi, M. (2016). Nietzsche on free will. *The Routledge Companion to Free Will*, ed. M. Griffiths, N. Levy, & K. Timpe. New York: Routledge.

Russell, P. (1992). Strawson's Way of Naturalizing Responsibility. *Ethics*, *102*(2), 287–302. http://www.jstor.org/stable/2381606

(2017) *The Limits of Free Will.* New York: Oxford University Press.

Scanlon, T. M. (1998). *What We Owe to Each Other.* Cambridge: Harvard University Press.

Schlosser, M. E. (2015). Manipulation and the zygote argument: Another reply. *The Journal of Ethics 19*(1), 73-84.

Schurger, A., Sitt, J. D., & Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences, 109*(42), E2904–E2913. https://doi.org/10.1073/pnas.1210467109

Shaw, Tamsin. (2007). *Nietzsche's Political Skepticism.* Princeton: Princeton University Press.

Strawson, P.F. (2005). *Scepticism and Naturalism: Some Varieties.* London: Methuen.

(2008). Freedom and resentment. *Free will and Reactive Attitudes*, 19-36, ed. Russell, P., & McKenna, M. New York: Routledge.

Slote, M. (2001). Moral responsibility without free will. *Genetics and Criminal Behavior*, 259–272, chapter 11, ed. D. Wasserman & R. Wachbroit Cambridge: Cambridge University Press.

Tomasello, M. (2016). *A Natural History of Morality.* Cambridge: Harvard University Press.

Van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon Press.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments.* Cambridge MA: Harvard University Press.

(2007). *Ressentiment*, value, and self-vindication: making sense of Nietzsche's slave revolt. *Nietzsche and Morality*, eds. B. Leiter and N. Sinhababu. Oxford: Oxford University Press.

Watson, G. (2004). *Agency and Answerability.* Oxford: Oxford University Press.

Wegner, D. M. (2002). *The Illusion of Conscious Will.* Cambridge, MA: The MIT Press.

Williams, B. (1981). Internal and external reasons. *Moral Luck.* Cambridge: Cambridge University Press.

(2002). *Truth and Truthfulness.* Princeton: Princeton University Press.

Zamosc, G. (2012), Sovereignty and Guilt in Nietzsche's *Genealogy*. *European Journal of Philosophy*, *20*(S1), E107-E142. https://doi.org/10.1111/j.1468-0378.2011.00484.x

Zeelenberg, M., Beattie, J., van der Plight, J., & de Vries, N. K. (1996). Consequences of regret aversion: Effects of expected feedback on risky decision making. *Organizational Behavior and Human Decision Processes, 65*(2), 148–158. https://doi.org/10.1006/obhd.1996.0013

(2009). On the consequences of mentally simulating future foregone outcomes: A regret regulation perspective. *Handbook of imagination and mental simulation*, K. D. Markman, W. M. P. Klein, & J. A. Suhr. Psychology Press.