DISCUSSING HIGHER-ORDER THEORIES OF CONSCIOUSNESS: BETWEEN HOT THEORY AND SELF-REPRESENTATIONALISM

ÇAĞLAN ÇINAR DİLEK

BOĞAZİÇİ UNIVERSITY

2025

DISCUSSING HIGHER-ORDER THEORIES OF CONSCIOUSNESS: BETWEEN HOT THEORY AND SELF-REPRESENTATIONALISM

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Philosophy

by

Çağlan Çınar Dilek

Boğaziçi University

2025

Discussing Higher-Order Theories of Consciousness:

Between HOT Theory and Self-Representationalism

The thesis of Çağlan Çınar Dilek

has been approved by:

Assoc. Prof. Lucas Thorpe (Thesis Advisor)	_____
Prof. Stephen Voss	_____
Assist. Prof. Sun Demirli	_____
Prof. David Rosenthal (External Member)	_____
Assoc. Prof Sinem Elkatip Hatipoğlu (External Member)	_____

June 2025

DECLARATION OF ORIGINALITY

I, Çağlan Çınar Dilek certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature……...

Date…….....

ABSTRACT

Discussing Higher-Order Theories of Consciousness:

Between HOT Theory and Self-Representationalism

This study examines whether higher-order thought (HOT) theory of consciousness should be adapted into a form of self-representationalism. I utilize the problem of misrepresentation as a tool to investigate the fundamental nature of HOT theory. The analysis reveals that HOT theory oscillates between the transitivity principle and the division of phenomenal labor. As the content of higher-order thought overdetermines conscious experience, the theory permits the occurrence of conscious experience in the absence of corresponding first-order mental states in cases of radical misrepresentation. Conversely, HOT theory is based on a relational structure involving first-order and higher-order mental states, wherein a first-order state becomes conscious if appropriately represented by a HOT. While the theory's allowance for misrepresentation is considered an advantage, the tension between two principles leads to additional problems. Most of the proposed modifications are independently motivated by attempts to account for an intrinsic notion of consciousness but it is not necessary to transform HOT theory into self-representationalism. Nevertheless, the relational structure of HOT theory remains, prompting ongoing efforts to develop alternative forms that preserve the existence of first-order states while maintaining the core principles of the HOT framework. Should the relational structure not be preserved, the theory risks collapsing into a variant of first-order thought theory. Exploring this possibility, I argue that integrating an account of a thick notion of self may better complement HOT theory compared to attempts that align it more closely with self-representational approaches.

ÖZET

Üst-Düzey Bilinç Teorilerini Tartışmak:
ÜDD Kuramı ve Öz-Temsiliyetçilik Arasında

Bu çalışma, üst-düzey düşünce (ÜDD) bilinç teorisinin bir tür öz-temsiliyetçilik biçimine uyarlanıp uyarlanamayacağını incelemektedir. ÜDD teorisinin temel doğasını araştırmak amacıyla yanlış temsil sorununu bir araç olarak kullanıyorum. Yapılan analiz, ÜDD teorisinin geçişlilik ilkesi ile fenomenal iş bölümü ilkesi arasında gidip geldiğini ortaya koymaktadır. Üst-düzey düşüncenin içeriği bilinçli deneyimi belirlediği için teori, radikal yanlış temsil vakalarında alt-düzey zihinsel durumlar olmadan da bilinçli deneyimin gerçekleşebilmesine izin verir. Diğer taraftan, ÜDD teorisi, alt-düzey bir zihinsel durumun, uygun bir üst-düzey düşünce tarafından temsil edildiğinde bilinçli hale geldiği, iki farklı düzeyde bulunan zihinsel durumlar arasındaki ilişkisel bir yapıya dayanır. Teorinin yanlış temsile izin vermesi bir avantaj olarak görülse de iki temel ilke arasındaki gerilim ek sorunlara yol açmaktadır. Önerilen çoğu değişiklik, bilinci içsel bir nitelik olarak açıklama girişimleriyle oluşturulmuştur, ancak ÜDD teorisini öz-temsiliyetçiliğe dönüştürmek gerekli değildir. Yine de ÜDD teorisinin ilişkisel yapısı problemli kalmaya devam etmekte ve bu da alt-düzey zihinsel durumların varlığını korurken ÜDD paradigmasının temel ilkelerini de uygulayan alternatif formların geliştirilmesini teşvik etmektedir. Eğer ilişkisel yapı korunmazsa teori, bir tür birinci-düzey düşünce teorisine dönüşme riski taşır. Bu olasılığı inceleyerek, ÜDD teorisini öz-temsiliyetçi yaklaşımlara daha fazla yaklaştırmaya çalışan girişimlere kıyasla, güçlü bir benlik kavramının entegre edilmesinin teoriyi daha iyi tamamlayabileceğini ileri sürüyorum.

CURRICULUM VITAE

NAME: Çağlan Çınar Dilek

DEGREES AWARDED

PhD in Philosophy, 2025, Boğaziçi University

MA in Philosophy, 2011, Boğaziçi University

BA in Philosophy, 2009, Boğaziçi University

AREAS OF SPECIAL INTEREST

Philosophy of mind, aesthetics, political philosophy

PROFESSIONAL EXPERIENCE

Instructor for an undergraduate-level course titled “Aesthetics”, Department of Fine Arts, Istanbul Technical University, 2015

Instructor for two undergraduate-level courses titled “Aesthetics”, Department of Fine Arts, Istanbul Technical University, 2012

PUBLICATIONS

Journal Articles

Dilek, Çağlan Ç. (2020). Zor problem ve yüksek düzey düşünce kuramı. *Phainomena*, *1*

Dilek, Çağlan Ç. (2017). Review of *I, me, mine* by Béatrice Longuenesse. *Phenomenological Reviews*, *3*, pp. 72

Conference Proceedings

Dilek, Çağlan Ç. (2025). Mental qualities without consciousness? Higher-order approach against one-level accounts. In *Boğaziçi University Philosophy Colloquim, Istanbul, Turkey.*

Dilek, Çağlan Ç. (2018). On the nature of representational relation in HOT theory: Extrinsicality, relationality and transitivity. In *CUNY Graduate Center Cognitive Science Speaker Series, New York, USA.*

Dilek, Çağlan Ç. (2017). Yüksek dereceden bilinç teorileri. In *İÜ Felsefe Topluluğu 29. Felsefe Kongresi, Istanbul, Turkey*

Dilek, Çağlan Ç. (2015). An ecological account of self-consciousness. In *Towards a Science of Consciousness, Helsinki, Finland*.

Dilek, Çağlan Ç. (2014). What is art: The main theories in contemporary philosophy of art. In *ITU Art Festival, Istanbul, Turkey.*

Dilek, Çağlan Ç. (2011). Can we understand self-consciousness through analyzing primitive self-awareness? Agency vs. self-consciousness: A discussion of phenomenological approaches towards self-consciousness. In *Towards a Science of Consciousness, Stockholm, Sweden*.

Other Publications

Dilek, Çağlan Ç. (2011). *Understanding Self-Consciousness* [Master's Thesis, Boğaziçi University].

ACKNOWLEDGMENTS

Completing this thesis has been a very long journey, and I am deeply grateful to those who have supported and guided me throughout this process.

First and foremost, I dedicate this work to the memory of my beloved mother Burçin. She always wished to see me complete my PhD thesis, and although she was unable to witness this milestone, her love, wisdom and unwavering belief in my potential have continuously inspired and motivated me.

My heartful gratitude goes to my dear wife, whose enduring support, patience, and sacrifices have made this achievement possible. Balancing the demanding care of our newborn child alongside my academic pursuits required immense strength and dedication, and I am forever grateful for her extraordinary support.

I extend my sincerest appreciation to my advisor, Assoc. Prof. Lucas Thorpe, for his invaluable support, insightful feedback and encouragement throughout my PhD thesis. His positive and enriching attitude made my studies joyful, just as it positively contributed to everyone around him.

I would also like to express my special gratitude to Prof. Stephen Voss, whose guidance and membership significantly sparked and cultivated my initial interest in self and consciousness. I am truly grateful for his invaluable contributions to my studies.

My thanks go as well to the esteemed jury members Assoc. Prof. Sinem Elkatip Hatipoğlu and Assist. Prof. Sun Demirli. Their positive attitudes, thoughtful comments and generous support greatly contributed to the refinement of my work.

Special thanks are owed to Prof. David Rosenthal for his invaluable insights, constructive feedback and continual engagement. His intellectual generosity in providing guidance and supportive attitude have been indispensable to my thesis research and overall scholarly development.

I am also deeply thankful to Prof. Howard Robinson, Prof. Philip Goff, Prof. Hanoch Ben-Yami, Assoc. Prof. David Weberman, Prof. Guenther Knoblich and Prof. Maria Kronfeldner, whose contributions during my PhD studies provided invaluable perspectives and significantly enriched my development as a researcher.

And finally, to my precious child, thank you for bringing endless happiness into my life. Your uplifting presence and boundless energy have filled each day with renewed purpose, motivated me to finish my thesis, and allowed me to experience the profound fulfillment of fatherhood.

to the memory of my beloved mother, who shaped the person I am today; to my loving wife, who sustains my present; and to my precious son, who illuminates my future.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

This thesis is an attempt to present a representational account of phenomenal consciousness that is higher-order in nature. My aim is to explain phenomenal consciousness, which I take as the combination of qualitative and subjective nature of our experience at a particular time. I believe that there is indeed something-it-is-like to experience a clear blue sky at a sunny spring morning, to drink a red Hungarian wine from Kékfrankos grapes, to listen to a captivating performance of Clair de Lune at a concert, to feel the warmth and softness of sand on a Mediterranean beach, to smell freshly baked cake emerging from the oven, to experience the position and movement of one's arms and muscles while working out at the gym, to feel excitement and happiness when reunited with loved ones after a prolonged absence, and to experience sadness or frustration due to the difficulties encountered while completing one's PhD thesis. These examples correspond respectively to the sensory states, proprioception, emotions, and moods. Such experiences frequently intertwine and seldom occur in isolation; crucially, there is always a subject undergoing these experiences.

I will refrain from making any assumptions regarding the nature of these experiences and I will not take a definitive stance on whether their explanation necessarily involves demonstrating how they belong to a subject. Following Kriegel, I distinguish between *qualitative character*, which refers to the whatness of an experience making a state phenomenally conscious it is, and *subjective character*, which refers to the for-meness of an experience, indicating there is something-it-is-like to experience a state at all (Kriegel, 2009, p.1). The explanandum is the unity of

qualitative and subjective character, termed by Kriegel as *phenomenal character*. This unity encapsulates what we signify by phenomenal consciousness and our expression that there is something-it-is-like, for instance, to see a red apple. While it might turn out that we can discuss phenomenal consciousness independently of its relation to the subject, I will argue that, strictly speaking, this is not viable. A theory of consciousness must account for both the qualitative and subjective dimensions of experience. Ultimately, the what-it-is-likeness of an experience exists only if this experience is for a subject. Therefore, the theory must either demonstrate that phenomenal consciousness is the unity of qualitative and subjective character, or that phenomenal consciousness predominantly consists of subjective character, while qualitative character can be explained independently of phenomenal consciousness.

The framework I will be working in will be higher-order and representational in nature, drawing primarily on David Rosenthal's higher-order thought (HOT) theory of consciousness. I think it is still one of the most promising theories of consciousness, particularly due to its giving separate explanations for consciousness and mental qualities. However, my approach is also influenced by phenomenological perspectives, emphasizing the minimal and non-representational form of subjectivity inherent in our experiences, as well as by self-representational theories, particularly those advocated by Uriah Kriegel, which prioritizes the self-representing subjective character present in our mental states. In this direction, this thesis aims to integrate key insights from these three approaches, examining whether some form of higher-order thought theory can be preserved as the most compelling account of phenomenal consciousness, while also incorporating phenomenological considerations and modifications proposed by self-representational theories.

I believe that HOT theory remains a strong alternative for explaining consciousness for several reasons: (1) It provides distinct explanations for consciousness and mentality, effectively addressing the challenges inherent in explaining consciousness within a materialistic world; (2) It offers an explanation in mentalistic terms, thereby avoiding reductionism and circumventing the need to solve the hard problem of consciousness; (3) It explains consciousness through unconscious mental states, which provides a naturalism-friendly framework and facilitates an understanding of how consciousness emerges from unconsciousness – a beneficial feature for those advocating physicalism; (4) It accurately captures the higher-order nature of consciousness and establishes clear distinctions among unconscious mental states, conscious mental states and introspection.

Higher-order (HO) theories are based on two simple principles: (a) The explanatory division of labor and (b) transitivity principle. According the former, we should adopt "a strategy of divide and conquer in the case of consciousness" (Rosenthal, 2005, p.1) to clearly distinguish our studies of mental qualities from those of consciousness itself. Mental states can have intentional or qualitative properties through which they are individuated, regardless of whether they are conscious. Consciousness, therefore, should be analyzed independently, in contrast to most philosophers who resist "considering consciousness apart from other features of mental functioning" (Rosenthal, 2005, p. 1). Importantly, adopting this strategy does not presuppose that mentality and consciousness are entirely distinct phenomena, nor is the strategy merely epistemological. Rather, the conclusion regarding their independence is justified on different grounds. Secondly, transitivity principle encapsulates the fact that mental states of which we are in no way conscious are unconscious mental states. From this fact arises a fundamental principle: “mental

states are conscious only if one is in some way conscious of them" (Rosenthal, 2005, p.4). While some regard this as a truism, others view it as a conceptual truth.

Although these two principles are considered foundational and positively regarded within the HOT paradigm, the explanatory division of labor in particular invites criticism. Firstly, opponents argue that detaching consciousness from mentality fails to adequately address phenomenal consciousness in terms of its qualitative, subjective nature. According to this criticism, HOT theory cannot explain phenomenal consciousness in its what-it-like-ness sense but only accounts for a technical, limited and weak sense of consciousness as higher-order consciousness. Accordingly, HOT theory explains away the hard problem of consciousness by decoupling consciousness from mental qualities and relying instead on an awareness-of relation through extrinsic HOTs. Critics maintain that this does not do justice to the intrinsic qualitative nature of consciousness, leaving phenomenal consciousness unexplained and the hard problem unresolved. Secondly, division of labor implies a problematic distinction between mental states, which have causal and functional roles, and consciousness, which is responsible only for appearance to the subject with minimal functional significance. Thirdly, the problem of misrepresentation indicates potential flaws in HOT theory's explanatory mechanism. Specifically, the precise role of first-order mental states in shaping conscious experience remains unclear, highlighting the concerning implication that higher-order thoughts alone determine conscious experience, regardless of whether the subject genuinely occupies those first-order mental states.

For the given reasons, later higher-order theorists strived to maintain the transitivity principle while adopting varied approaches to detach consciousness from first-order qualities. This is done mostly by formulating a type of self-

representational theory characterized by a complex mental state composed of lower and higher-order mental states as their parts. Accordingly, a complex mental state becomes conscious when its lower-order part is represented by its higher-order part. This approach preserves the primary strengths of HOT theory: Consciousness is explained as state consciousness via the relation of being represented as outlined by the transitivity principle, distinguishing it from first-order representational theories that explain state consciousness in terms of a state's representation of external objects. Additionally, consciousness becomes an intrinsic property of the whole complex state, while it is explained relationally at the same time.

I believe that these problems are significant and require further elaboration and additional justification to demonstrate that a notion of consciousness, when explained independently of mental qualities, can adequately capture the notion of what-it-is-likeness in the relevant sense. However, it seems that the majority of criticisms stem from disagreements about what constitutes phenomenal consciousness and what-it-is-likeness. Higher-order representational theories conceive of consciousness as an appearance to a subject, while they explain our mental life through intentional and qualitative features without involving the notion of subjectivity, thereby omitting what-it-is-likeness. For higher-order thought theories, therefore, the notion of what-it-is-likeness only makes sense if it is for a subject. It is preferable to say that there is nothing-it-is-like to see red if that perceptual state is unconscious, because it is not appropriately represented at a higher-order level. On the other hand, opponents of higher-order theories seek to explain phenomenal consciousness through explaining the qualitative experiences at the first-order level. For them, there is something it is like to see red without the necessity of that state represented or accessed at a higher-level. Hence, consciousness

is explained intrinsically through qualitative properties, and what-it-is-likeness resides at this first-order level. It is a further question whether there is subjectivity at this level or whether the concept of subjectivity, as a self, is a separate notion independent of phenomenal consciousness. This latter position tends to be endorsed by theorists who regard the problem of consciousness as hard or special, treating phenomenal consciousness as a unity of mentality and consciousness, without distinguishing between qualitative and subjective characters (Kriegel, 2009).

In this thesis, I will defend higher-order thought theory and argue against transforming it into a self-representational theory, particularly if the motivation behind such transformation is to render consciousness an intrinsic property of the whole mental state or to accommodate the intuition that consciousness is reflexive. However, I maintain that certain modifications are necessary to develop a stronger HOT theory. A careful examination of HOT theory often gives rise to the intuition that changes are necessary; however, upon deeper exploration of its theoretical framework, it becomes less clear exactly which aspects should be adjusted or refined.

Firstly, while pursing division of labor as an epistemological method, I have suspicions about whether a higher-order thought can give us all about what-it-is-likeness. Indeed, most criticisms target HOT Theorists' approach of depriving mental states of their what-it-is-likeness, and I believe there is some merit to these objections. Therefore, we must explain the phenomenal character present in mental states that are not properly conscious (in higher-order sense). Analyzing the role of self within the higher-order thought theory could help us to resolve many disagreements between HOT theories and their critics. Specifically, it is essential to clarify how the self is closely related to higher-order thoughts and how this connection underpins our consciousness. In doing so, the higher-order framework

must incorporate an account of the pre-reflective, minimal subjectivity present in our mental life, a feature emphasized in non-representational phenomenological approaches. While rejecting first-order representationalism that explains phenomenal consciousness through qualitative properties, HOT theories can find unexpected allies in the phenomenological framework – of course to some degree. Consequently, HOT theory need not to be modified into a self-representational theory. However, it requires a better articulation of self, alongside an explicit recognition of the minimal subjectivity already present in our mental lives, without explicit consciousness.

1.1 The structure of the thesis

In Chapter 1, I will introduce the central problem and situate it within a broader philosophical context by relating it to other relevant issues in the study of consciousness. I will present the basic terminology, carefully clarifying the notions of what-it-is-likeness and the various conceptions of consciousness, since verbal disagreements frequently obscure philosophical debates about consciousness. Subsequently, I will outline the main features of classical HOT theory as developed by David Rosenthal and provide a comparative analysis of his theory with other representational accounts.

In Chapter 2, I will focus on the problem of misrepresentation. HOT theory has frequently been scrutinized due to its handling of cases involving illusion and hallucination. Although such cases are not unique to higher-order theories, the possibility of such unsuccessful cases of perception is often regarded as vulnerability within the higher-order paradigm. This vulnerability arises because consciousness, according to HOT theory, is explained extrinsically through a representational relation between two mental states, which can relatively easily malfunction.

Additionally, the nearly exclusive reliance on higher-order thoughts by HOT theory when explaining the nature of conscious experience generates additional fundamental issues. My analysis will examine these interconnected problems to uncover the core assumptions that lead to theoretical disagreements within the HOT framework.

In Chapter 3, I will argue that HOT theory does not require modification into a form of self-representationalism. To substantiate this position, I will analyze various forms of self-representationalism. First, I will examine historical examples of pure self-referential theories. Following that, I will assess Rocco Gennaro's wide intrinsicality view as a form of intrinsic HOT theory. Finally, I will conclude with an analysis of Kriegel's self-representational theory.

In Chapter 4, I will first focus on the question whether consciousness is independent from mentality. In HOT theory, the answer has two parts: The first part is related to the hard problem of consciousness, while the second part aims to offer a positive account of mental qualities independent of consciousness. The first part deals with classical problems such as the *knowledge argument*, *inverted* (and *absent*) *qualia*, and Chalmers' *philosophical zombies*. It is demonstrated that these concepts are based on a wrong intuition, which is often regarded as pre-theoretical, but actually relies on foundational theoretical assumptions. HOT theory advances both the weaker epistemological claim that we need to investigate consciousness independent of our mental life and the stronger ontological claim that consciousness is independent of mentality. In the second part, Rosenthal uses quality-space theory to explain how mental qualities function, why we should understand such qualities as genuinely qualitative, and why our asymmetric nature of our perceptual system precludes the inversion of our mental qualities as envisioned in classical thought experiments.

Finally, I will discuss a persistent issue within HOT theory arising from the discussion of misrepresentation. The main problem is not that consciousness needs to be explained as an intrinsic quality of first-order mental states; rather, HOT theory continuously deals with a tension between division of phenomenal labor and transitivity principle. Although HOT theory is based on the transitivity principle and a relational structure between first-order and higher-order states, Brown's distinction between relational and non-relational HOT theories proves insightful. The most productive modification to HOT theory would involve integrating a strong notion of self. Consequently, consciousness might be explained through a non-relational HOT account, prioritizing inner-awareness as its foundational principle.

1.2 HOT theory of consciousness

Phenomenally conscious states are mental states with subjective dimension. In Nagelian terms (1974), there is something it is like to undergo a conscious experience that has a distinctive qualitative feel. Explaining what makes a creature conscious is not crucial for this inquiry since a conscious being can have both conscious and unconscious mental states. *Creature consciousness* refers to a state of being wakeful and mentally responsive (Rosenthal, 2005, p. 46). It is more important for us to explain how some states are conscious while others are not. The subject of our investigation, as well as the focus of the hard problem of consciousness, is *state consciousness*. How do certain mental states experienced by a conscious subject become identifiable as conscious? How are these states presented or made available to the subject? Who precisely is aware of these experiences, and in what particular manner?

Block's (1995) distinction between *phenomenal consciousness* and *access consciousness* plays an important role in these discussions: "A state is access-conscious (A-conscious) if … a representation of its content is (1) … poised for use as a premise in reasoning, (2) poised for rational control of action, and (3) poised for rational control of speech" (p. 231). On the other hand, Block defines phenomenal consciousness as experience. Accordingly, there is no noncircular way to explain what phenomenal consciousness is and this is not a problem considering the "history of reductive definitions in philosophy" (Block, 1995, p. 230). The experiential properties of sensations, feelings, perceptions, thoughts, desires and emotions constitute phenomenal-consciousness properties and "the totality of the experiential properties of a state are what it is like to have it" (Block, 1995, p. 230).

According to Block, access consciousness (A-consciousness) does not guarantee phenomenal consciousness (P-consciousness); a subject's awareness of a mental state does not necessarily make it phenomenally conscious. Conversely, a state can be phenomenally conscious without being access-conscious, implying there would still be something-it-is-like to experience that state. In this context, phenomenal consciousness is the object of hard problem of consciousness (Chalmers, 1997) and focus of questions such as "Why do certain physical brain processes result in the subjective experience of color, rather than experiences of sound or no experiences whatsoever?" (Weisberg, 2014, p. 1).

Rosenthal utilizes the distinction but is entirely critical of Block's framework. His study also targets phenomenal consciousness "rather than access-consciousness or self-consciousness that has seemed such a scientific mystery" (Block, 1995, p. 230). However, Rosenthal argues that the idea that no noncircular explanation of phenomenal consciousness can be given is based on a theoretical assumption,

ultimately resulting in an uninformative explanation. As we will explore throughout this thesis, Rosenthal explains phenomenal consciousness through subject's awareness of a mental state. Nevertheless, he rejects the idea that there can be an awareness relation without phenomenal consciousness. It is crucial to recognize both the similarities and differences between Rosenthal's awareness relation and Block's notion of access-consciousness: Whereas Block seeks to highlight the uniqueness of phenomenal consciousness by distinguishing it from a mere cognitive and non-experiential access-consciousness, Rosenthal employs a cognitive relation to explain phenomenal consciousness in a naturalistic framework.

Nagel questions our ability to answer the question what it is like to be a bat. Can I, as a human individual named Çağlan, ever know, what it is like to be a bat? I inherently lack the necessary sensory and cognitive capacities to fully comprehend the distinctive perceptual experiences of a bat. Is this limitation merely due to my species-specific biology and cognitive framework? Not exactly. Accordingly, there is a deeper, more fundamental aspect of conscious experience itself that prevents me from knowing what-it-is-like to be a bat; namely, that conscious experience has a subjective point of view, which can only be fully known through direct experience. Such knowledge requires one to be the subject undergoing that experience. Frank Jackson (1982) explores a similar idea through his thought experiment, *Mary the color scientist.* Could Mary, who has spent her entire life in a black-and-white room, and thus being exposed to only black and white objects, truly know what it is like to see a red apple? Assuming she has acquired all the physical facts about color vision, does she acquire new knowledge when she first leaves her room and see a red apple? If we answer affirmatively, it suggests that there are some facts about mental properties accessible only through consciousness, pointing to a distinction between

third-person facts and first-person facts. The subjective knowledge possessed by the bat and the novel knowledge Mary acquires upon first seeing a red object both correspond to the what-it-is-likeness of their respective experiences. These are mental states with particular contents giving rise to phenomenal consciousness. I cannot fully grasp the subjective experience of a bat from a third-person perspective. Similarly, Mary cannot acquire the phenomenal quality experiencing redness solely by studying physical facts within her black-and-white room.

Chalmers distinguishes between the *easy* and *hard problems of consciousness*. According to Chalmers (1997), easy problems encompass explaining mechanisms such as attention, behavioral control, the distinction between wakefulness and sleep, our capacity to report mental states, information organization, object categorization, and our capacity to react (pp. xii-xiii). These problems are labeled easy not because they require minimal effort or time to give satisfying explanations. Neuroscience, psychology, and cognitive sciences may take over a century to fully explain these phenomena, if they ever do. However, for Chalmers, these problems are easy because there are no fundamental theoretical obstacles preventing us from solving them. These problems relate to cognitive functions and can typically be addressed by describing mechanisms, either at a detailed neural level or at a more abstract computational level. For example, explaining internal access involves describing how a system processes internal states and the specific type of information utilized to manage such states. In contrast, the hard problem transcends functional explanations and the performance conditions of explained functions. Here, the question shifts from how certain cognitive functions are performed to why they are accompanied by conscious experiences with phenomenally rich qualities. Chalmers specifically questions why cognitive functions do not occur without

subjective experiences – why these cognitive tasks are not carried out in darkness. The absence of a satisfactory explanation for this question constitutes the explanatory gap (Levine, 1983), highlighting the profound division between functional states and conscious experience. Accordingly, it is not possible to identify physical or psychological explanations with conscious mental states. Consequently, phenomenal consciousness will remain mysterious and inexplicable.

It is methodologically problematic to define consciousness as "the mysterious, unknowable core of human experience" (Weisberg, 2014, p.1), because such a definition inherently resists any explanation and remains mysterious. Weisberg (2014) proposes instead using the working definition "whatever causes our verbal reports about how we feel" (p.2). I would argue that mysteriousness is indeed an important aspect of this phenomenon, suggesting a definition along the lines of whatever causes the mystery surrounding consciousness and resists explanations presented by the solutions to the easy problem of consciousness.[1] With this apprach, we can arrive at an explanation for consciousness that is neither mysterious nor inexplicable. It also enables us to question the existence of a distinct hard problem and challenge the notion that thought experiments reveal knowledge accessible solely through consiousness. Rosenthal (2010) argues that these position stem from unjustified theoretical assumptions:

> The ostensibly pretheoretic intuitions … all rest on the view, seldom explicitly articulated, that consciousness is our only source of knowledge about mental qualities. But that's not a pretheoretic intuition, but a theoretical claim about mental qualities, a claim that one might well contest. So, the intuitions about quality inversion, zombies, the explanatory gap, and the hard problem, though at first sight they may seem independent of theory, all rest on a theoretical claim about mental qualities. Unless there is no alternative to

[1] Kriegel (2009) follows such a route: "The term 'consciousness' is, of course, an ordinary, everyday term to be understood intuitively. We can then use this intuitive understanding in fixing the reference of the technical term 'phenomenal consciousness' with the following rigidified definite description: the property F, such that, in the actual world, F is responsible for the mystery of consciousness.'' (p. 3)

> theorizing about mental qualities in that way, those alleged intuitions cannot serve as data that any satisfactory theory must accommodate. (p. 372)

We can seek to explain phenomenal consciousness and the notion of what-it-is-likeness without necessarily endorsing the view that all mental states are inherently conscious. Additionally, we need not accept the claim that a functional explanation of consciousness is inherently impossible. Higher-order representational theories explain consciousness through the representational relation between two mental states. Such theories align comfortably with physicalism, but they do not inherently presuppose its necessity, as their explanations operate at a psychological level making use of mental terms. Moreover, their primary aim is not to defend functionalism per se, a debate typically centered around thought experiments by Ned Block, Frank Jackson, and David Chalmers. David Rosenthal explicitly challenges the idea that these theories rely on pre-theoretical intuitions and argues that folk-psychology is disguised under a particular understanding of mentality which presents itself as non-theoretical. He maintains that multiple conceptions of mentality exist, and he endeavors to provide an informative explanation by distinguishing mentality and consciousness. This distinction is precisely what makes higher-order theories fruitful and promising: If successful, HOT theory can effectively distinguish between physical, mental and conscious properties.

Consciousness-based views maintain that all mental states are inherently conscious, since consciousness is considered an intrinsic property of mental states. Rosenthal criticizes this approach as uninformative and misleading, arguing instead for distinct theories of mental qualities and consciousness due to explanatory and conceptual considerations. HOT theories achieve this by explaining mental qualities through their perceptual roles, and consciousness through higher-order thoughts. These theoretical components are methodologically independent; thus, HOT theory

is not merely the aggregate of Rosenthal's views on mental life but it is explicitly focused on the explanation of phenomenal consciousness.

1.3 Terminology for higher-order representational theories

Representational theories are based on four distinctions: (1) Creature consciousness and state consciousness; (2) Intransitive consciousness and transitive consciousness; (3) Phenomenal consciousness and access consciousness; (4) Subjective character and qualitative character (Kriegel, 2009; Rosenthal, 2005).

A creature can be characterized as conscious or unconscious. For example, we might describe Michael Schumacher as unconscious if he is currently in coma. Similarly, an individual is generally considered unconscious while sleeping. Additionally, mental states themselves can be conscious or unconscious. A conscious creature can have both conscious and unconscious mental states, and Rosenthal even argues that unconscious creatures may have conscious mental states. Our primary interest lies in state consciousness, specifically in distinguishing between conscious and unconscious mental states.[2]

Burçin can be described as being *intransitively conscious* because she is awake. At the same time, she can be said to be *transitively conscious* of the sunlight entering through her window. The transitive/intransitive distinction applies both to creatures and mental states. If I see the red wine in front of me, I am transitively conscious of the red wine, but I also have a transitively conscious mental state with the content that I am seeing the red wine. In this case, the mental state can also be characterized as intransitively conscious (M is conscious = M is a conscious state).

[2] We can distinguish *species consciousness* from *creature consciousness* as a different form of intransitive creature consciousness. In this discussion, I will follow Kriegel (2009) and accept that "the truthmaker of 'orangutans are conscious' consists in many orangutans instantiating creature consciousness, not of the species instantiating a new property." (p. 29)

These distinctions are crucial to understand how higher-order theories operate. Can a mental state be transitively conscious? Higher-order thought theory explains phenomenal consciousness as intransitive state consciousness via transitive consciousness. It remains debatable whether the transitive consciousness at issue is best understood as a property of the creature or the state. Arguing for the former considers it a transitive property of the subject, while arguing for the latter views it as a transitive property of the higher-order state.

In HOT theory, the intransitive usage of state consciousness serves as the primary object of investigation, while the conscious-making process is explained through the transitive conscious-of relation. Kriegel argues that intransitive state consciousness is prior to the other three forms of consciousness, but this stance diverges from Rosenthal's perspective.[3] According to HOT theory's classical formulation, a mental state becomes conscious in virtue of being represented by an appropriate higher-order thought. Hence, one may say that a represented first-order mental state becomes intransitively conscious through being represented by an appropriate HOT. Alternatively, this process can be articulated by stating that the subject becomes intransitively conscious by virtue of having an appropriate HOT. If consciousness is regarded as a property of the state, we explain intransitive consciousness of a first-order mental state by appealing to the transitive consciousness of a higher-order state whose object is the first-order state. Conversely, if consciousness is treated as a property of the subject, the subject's intransitive consciousness can be elucidated through the subject's transitive consciousness of their own experience, which, in turn, can be explained through the

[3] Kriegel takes one more step and carries the distinction between transitive and intransitive to self-consciousness. He argues that intransitive self-consciousness is prior to all the other kinds of consciousness and self-consciousness. (Kriegel, 2004)

transitive state-consciousness of the mental state subject has. Consequently, conscious states can be explained as "states of which the subject is aware, or states of which the subject is creature-conscious" (Carruthers & Gennaro, 2020, Sec. 2). Kriegel's view on the primacy of state-consciousness suggests a form of consciousness inherent within mental state. Although Rosenthal indeed explains intransitive state consciousness, he explicitly rejects this intrinsic conception. Nevertheless, Rosenthal would not characterize his own position as specifically advocating for the primacy of transitive consciousness either.

Relating mental states to creature consciousness is partly a distinct issue within classical HOT theory, as state consciousness and creature consciousness are not entirely interdependent. According to Rosenthal, a property of a state differs from the property of a creature; therefore, state-consciousness is distinct from creature-consciousness. Even though these notions are not always explicitly differentiated, we should bear in mind that they represent distinct categories, and knowledge about one does not necessarily provide much insight into the other (Rosenthal, 1990, p. 730). Creature consciousness pertains more directly to being awake, and there is no particular mystery surrounding it – at most, the mystery is akin to that previously associated with vitalism and the question of how life could emerge in a purely physical world. Just as the concept of life eventually lost its enigmatic nature by becoming well-understood, the understanding of creature consciousness similarly became largely explicable, no longer shrouded in mystery. Given these considerations, it is important to recognize the distinctions among various higher-order thought theories and Rosenthal's classical formulation. The critical point to emphasize is that transitive state consciousness serves as the primary explanatory tool for accounting consciousness in the intransitive sense.

Phenomenal consciousness refers to what-it-is-likeness of an experience, whereas access consciousness refers to an individual's access to their mental states. There are various functionally definable forms of access-consciousness. The main question is whether phenomenal consciousness can be reductively explained in functional or representational terms like access consciousness (Carruthers & Gennaro, 2020). There is general consensus that not every instance of access-consciousness is phenomenally conscious. However, within higher-order frameworks, there is a tendency to maintain that all phenomenally conscious states are also access-conscious, in the sense that the subject is aware of their states by being able to represent them.

Kriegel introduces a distinction between *subjective character* and *qualitative character* to distinguish his self-representational theory from classical HOT theory. Qualitative character refers to the mental qualities such as redness and sharpness, whereas subjective character pertains to the for-meness, or the subjectivity, of a mental state. Phenomenal character encompasses the overall what-it-is-likeness of an experience and constitutes the object of our inquiry as phenomenal consciousness. Theories vary based on how they interpret phenomenal character in terms of qualitative and subjective characters. According to HOT theory, it is possible to have an unconscious mental quality, indicating the existence of qualitative character without subjective character. However, HOT theory defines phenomenal character primarily through subjective character, because the HOT theorist argues that there cannot be a conscious state of which one is not aware. Consequently, subjective character overdetermines phenomenal character (Kriegel, 2009, Ch. 2).

Representational theories seek to explain consciousness in representationalist terms to align with a naturalistic and physicalist framework. First-order

representationalism explains consciousness “in terms of world-directed intentional states” (Gennaro, 2012, p. 13). Proponents argue that a mental state is “conscious if and only if it represents the right way” (Kriegel &Williford, 2006, p. 1). According to Tye, the phenomenal properties of an experience constitute its representational content (Gennaro, 2012, p. 13). The notion of *transparency of experience* is used to support this view: If I am looking at my brown desk and briefly reflect on my own perception, I become aware only of the brownness of the desk itself. Thus, experience is considered transparent, rendering unnecessary any secondary mental state directed toward first-order experience to introduce an additional second-order property as the content of conscious experience. Kriegel (2007) formulates the argument for representationalism as follows:

> 1. The only introspectively accessible properties of conscious experience are its representational properties;
> 2. The phenomenal character of conscious experience is given by its introspectively accessible properties; therefore,
> 3. The phenomenal character of conscious experience is given by its representational properties. (p. 44)

The first premise serves as the *thesis of transparency*, from which representationalism follows (Premise 3) through a conceptual truth (Premise 2). Higher-order representational theories argue that a mental state is conscious if and only if “it is represented in the right way” (Kriegel and Williford, 2006, p.1). Their main tool is *transitivity principle*: “mental states are conscious only if one is in some way conscious of them” (Rosenthal, 2005, p. 4). This principle is sometimes expressed negatively: A mental state, of which one is in no way conscious, is not conscious. This definition strongly opposes *consciousness-based views* that suggest all mental states are conscious. Kriegel (2007) constructs the main argument of higher-order representationalism as follows:

> 4. A mental state M of subject S is conscious when, and only when, S is aware of M in the appropriate way;
> 5. Awareness of X requires mental representation of X; therefore,
> 6. M is conscious when, and only when, S has a mental state M*, such that M* represents M in the appropriate way. (p. 48)

The second premise (Premise 5) is widely accepted across most representational theories of consciousness, while the first premise (Premise 4) serves as a defining characteristic of higher-order theories and forms the basis for most of the discussions within the literature. This premise is formulated in various ways by Block, Carruthers and Kriegel:

> A phenomenally conscious mental state is a mental state (of a certain sort…) that either is, or is disposed to be, the object of a higher-order representation of a certain sort. (Carruthers & Gennaro, 2011, Sec. 2)

> A mental state is conscious if and only if the state is the object of a certain kind of representation arrived at non-inferentially. (Block, 2011, p. 421)

> A mental state M of a subject S is conscious iff S has a mental state M*, such that (i) M* is an appropriate representation of M, and (ii) there is no constitutive relation between M and M*. (Kriegel, 2006, p. 143)

Carruthers keeps higher-order representation vague, allowing it to be either a perception or thought, and either *dispositional* or *occurrent*. Block emphasizes non-inferentiality as the factor that determines what makes a representation direct. Meanwhile, Kriegel focuses on the contingent nature of the relation between two parts of the representation to distinguish higher-order theories from self-representational accounts, where this relation is viewed as necessary and constitutive.

1.4 Main features of HOT theory

Mental states are either in sensory or intentional character. In this direction, it makes sense to look to these properties to explain consciousness. Higher-order theories mainly differ on whether they explain consciousness through emphasizing sensory qualities or intentional qualities – in terms of how they see the nature of the higher-

order state. Theories coming from the tradition of Locke and Armstrong, which view the meta-state as a monitoring mechanism and thus quasi-perceptual in nature, are called *higher-order perception* theories. On the other hand, Rosenthal defends the view that the meta-state is a thought, resulting in what are known as higher-order thought theories.

Higher-order perception theories attempt to associate consciousness with a distinct sensory modality, analogous to taste or vision, to explain its rich content and experiential responses. However, as Rosenthal (1997) argues, "mental states do not exemplify a single range of properties" like a sensory modality does (p. 740). Various conscious states do not share the similarities in the same manner that experiences within a single modality do. For instance, a visual experience of a red apple significantly resembles the visual experience of a green apple, but there is no comparable similarity between a conscious visual experience and a conscious auditory experience in this aspect.

Similar issues do not arise when considering the alternative that the higher-order state is a thought. "Thinking is not specific to a particular range of objects, it involves no dedicated organ in the way sense modalities do … Nor does having thoughts about things involve any characteristic qualities" (Rosenthal, 1997, p. 741). Rosenthal finds it problematic for a higher-order state to be perceptual, as this would require an additional mental quality distinct from the first-order perceptual quality. If the physical object is the object of perception, then the perceptual property of the state would be the object of higher-order perception. However, there is no perceptual quality uniquely associated with consciousness, unlike the distinctive qualities of sensory experiences. If such qualities existed at the first-order state, this would transform the higher-order perception into a single-level account, making

consciousness intrinsic to the first-order perceptual state – precisely the scenario higher-order thought theories aim to avoid. Rosenthal argues that we are conscious of things by either perceiving or having a thought about them as being present. Thus, if the higher-order awareness lacks perceptual character, consciousness of a mental state should derive from our having thoughts about those states (Rosenthal, 2009, p. 243).

It is not sufficient for a mental state to be accompanied by a HOT to become conscious. The higher-order representation must occur in an appropriate manner: It must be occurrent (rather than dispositional), subjectively noninferential and subjectively unmediated. One cannot simply direct a HOT toward a first-order state at will to make it conscious. Rather, it must subjectively seem to the individual that their HOTs are direct and not mediated. Whether these HOTs are really unmediated is a different issue, as numerous mental processes involve mediation through various cognitive processes that remain unnoticed by the subject. Indeed, if there were genuinely no mediation, one would hold a consciousness-based view. For our purposes, it suffices that our having HOTs occur automatically and non-inferentially.

CHAPTER 2

THE PROBLEM OF MISREPRESENTATION

HOT Theories face criticism mainly due to the issue of misrepresentation. One challenge is that the theory allows such errors. However, the possibility of misrepresentation is not inherently problematic, as individuals often experience forms of misrepresentation in daily life. For instance, one might mistake an apple for an orange or perceive a green apple when it is actually red. Both cases are examples of illusion (*ordinary misrepresentation*). There are also cases where individuals perceive objects without any sensory input, known as hallucination or *radical misrepresentation* (*targetless misrepresentation*) in the higher-order framework. I agree with Rosenthal and consider both types of misrepresentation as part of the same problem, although the radical misrepresentation is more severe and poses greater challenges for HOT theory. The primary issue lies in how HOT theory addresses these cases without either (a) falling into inconsistency, or (b) sacrificing its core principles and evolving into a different theory. Moreover, the problem of misrepresentation has broader implications: It highlights controversial aspects of HOT theory, which have been subject to extensive critique by its opponents.

2.1 The problem of misrepresentation

The criticisms regarding inconsistencies concentrate on the conflict between the necessary and sufficient conditions of the theory. On one hand, higher-order thoughts play a dominant role in determining the nature of our conscious experience; they are sufficient for consciousness to occur and their content determines the nature of our experience. In cases of hallucination, there is a conscious experience without a

corresponding first-order state, which calls into question the role of first-order states and complicates the understanding of veridical perception. Moreover, discussing a consciously non-existent first-order state appears counterintuitive. On the other hand, if the contents of both first-order representation (FOR) and higher-order thought (HOT) are jointly necessary for consciousness, and if the phenomenal quality is determined in a bottom-up manner from lower-order perception, it is uncertain whether we ultimately adhere to a higher-order theory. The theory risks evolving into a first-order or self-representational theory. For these reasons, addressing criticisms related to the problem of misrepresentation and examining the higher-order response is crucial for comprehending the core features of HOT theory and evaluating its efficacy in explaining phenomenal consciousness.

Firstly**,** I will present the main formulations of the criticism based on misrepresentation. Then I will discuss the problems that HOT theory faces in relation to the misrepresentation problem, not only in hallucination and illusion but in successful perception and thinking in general. These problematic features of HOT theory will be examined under the following headings: (a) Division of phenomenal labor, (b) overdetermination of HOT, (c) the risk of collapse into a first-order theory, (d) the possibility of non-existent but conscious first-order states, and (e) a particular HO reading of what-it-is-likeness. Lastly, I will address how David Rosenthal and other higher-order theorists respond to these criticisms.

The issue of misrepresentation will be pertinent to the discussions in subsequent chapters, especially in relation to self-representational theories addressed in Chapter 3, because one of the motivations behind constructing a self-representational theory is to prevent the possibility of misrepresentation.

2.1.1 Formulating the problem of misrepresentation

Most of the criticisms in the literature (Block, 2011; Levine, 2001) base their position on the way Karen Neander (1998) formulates the problem. Neander attacks Lycan's higher-order perception (HOP) theory but the difference between HOP and HOT theories on this problem is negligible. Neander (1998) presents three possible scenarios to consider:

> (a) I have a sensory representation of something (existent or inexistent) as green. And I represent myself as representing something as green. (b) I have a sensory representation of something (existent or inexistent) as red. But I represent myself as representing something as green. (c) I have no sensory representation at all. But I represent myself as representing something as green. (p. 420)

First case (a) is a successful perception according to a higher-order (HO) theory. There is a first-order perception that "I see x with the property green" and a second-order representation of the FOR so that I am conscious that I see a green x.[4] Second situation (b) is an example of illusion - there is a first-order perception as of a red object, which is misrepresented as being green. This is the paradigm case of misrepresentation. The third one (c) is an example of hallucination – there is no first-order perception, while the subject has the conscious experience as of a green object. This case is called targetless misrepresentation or radical misrepresentation in the literature to distinguish it from normal misrepresentation. This issue constitutes the primary challenge for HOT theory and engaging with it further illuminates the nature of HOT theory. However, many opponents as well as defenders of HOT theory do not find it necessary to distinguish between normal and radical misrepresentation. Neander identifies different strategies that higher-order theorists might adopt in cases (b) and (c). Similarly, most current discussions focus on whether there is any

[4]According to higher-order perception theory, second-order representation is perceptual in nature and for HOT theory it is a thought. So, for HOT theory, the content of the higher-order thought would be that *I think that I see x with the property green*.

plausible route that a higher-order theory can adapt, either by maintaining or revising some of its core features.

Neander emphasizes two distinct features within Lycan's theory that are crucial when addressing the issue: "(A) Sensory representations determine the quality of a sensation, self-monitoring representations determine if the sensation is conscious. (B) The conscious experience of a sensation is (is determined by) the second-order representing of the first-order sensory representation" (Neander, 1998, pp. 421-422). Accordingly, the first feature aims to give an account of qualia, while the second one seeks to explain consciousness. This distinction leads to different interpretations of misrepresentation cases b and c. Feature A incorporates first-order perceptions by aiming to explain qualia at the primary level, while feature B assigns the determining role to higher-order representation (monitoring mechanism), thus elucidating consciousness through a higher-order representational mechanism.[5] In this context, with feature A, Lycan deals with the problem while trying to explain mental qualities and consciousness in different ways, and this is what Neander calls division of the phenomenal labor. This approach generates ambiguities regarding the precise nature of mental features. If they are phenomenal, what is added with self-monitoring? It is also problematic what kind of role they play when a mental state is conscious with a self-monitoring mechanism. On the other hand, feature B dissolves the division of phenomenal labor by attributing all consciousness-determining role exclusively to the higher-order state. Consequently, there is a conflict between two features of Lycan's theory.

[5] One should be careful not to utilize the term “qualia” in such a discussion because it has a usage associated with what-it-is-likeness. In this respect, it situates consciousness within the mental quality, while HOT theorists mainly aim to give different explanations for mental quality and consciousness, where mental qualities can be unconscious. Thus, it is more neutral to talk about mental qualities in place of qualia.

Block (2011) examines Rosenthal's theory as oscillating between a *modest* and *ambitious* version, similar to the distinction Neander identifies in Lycan's theory. The modest version aims to explain consciousness in the technical sense of *higher-order consciousness*, while acknowledging the presence of a separate notion of what-it-is-likeness or phenomenal consciousness at the first-order level. In this framework, "an unconscious pain might nonetheless be conscious in the what-it-is-likeness sense of the term" (Block, 2011, p. 422). Conversely, the ambitious higher-order theory attempts to explain consciousness, including what-it-is-likeness, by asserting that what-it-is-likeness does not exist without higher-order representation. Block argues that modest version is not interesting and does not address the problem of consciousness as understood in the literature. He contends that Rosenthal's theory should be considered in its ambitious form, which must then provide a satisfactory explanation for how there can be no what-it-is-likeness to experience pain in the absence of a relevant higher-order representation.

Returning to Neander's scenarios, a HOT theorist could explain scenario b in several ways. There might be (i) no conscious experience, (ii) a conscious experience of something red, or (iii) a conscious experience of something green. According to higher-order thought theory, for a mental state to be conscious, it should be represented by a higher-order thought in an *appropriate* (occurrent, non-inferential and subjectively direct) way. If we follow Lycan's second feature (B) or an ambitious HOT theory, we will choose the option b that there is a conscious experience of red. One is conscious of the way HOT represents and the content of first-order representation does not determine the nature of one's conscious experience. However, if we follow Lycan's first feature (A) or a modest HOT theory, the outcome is less clear and requires further explanation through additional theoretical

premises. Conflicting first-order and second-order representations lead to either no conscious experience (i), or to one content overdetermining one's conscious experience (ii or iii). A modest HOT theory, which explains consciousness of a first-order mental state through its appropriate representation, does not clearly address what occurs when that first-order state is misrepresented. Similar issues arise in radical misrepresentation (c). An ambitious HOT theory would conclude there is a conscious experience of red object, because the theory takes the presence of HOT as sufficient for conscious experience to occur. This is problematic because first-order representation (FOR) does not seem to play a role on our conscious experience at all.[6] Conversely, a modest HOT theory would not be conclusive about the nature of conscious experience in such a case: It considers both FOR and HOT as necessary and jointly sufficient for consciousness to occur. As there is no first-order perception in case (c), it is not clear for modest theory what the resulting experience would be.

Block formulates the problem in the following way: The theory supplies a necessary and sufficient condition for a conscious episode. An episode is conscious at t if and only if it is the object of an assertoric higher order thought at t, arrived at non-inferentially. The sufficient condition dictates that this thought at t is sufficient for a conscious episode at t. By the necessary condition, that conscious episode at t is the object of a simultaneous higher order thought. In the example, there is only one higher order thought at t, and we can assume it is not self-referential. So, there is no conscious episode at t after all. Thus, the sufficient condition and the necessary

[6] Normally, first-order states play a causal role in the occurrence of HOTs, and thus contribute to the generation of consciousness. However, there is no causal requirement that a particular first-order state must cause the occurrence of a corresponding HOT, nor that the intentional object of a HOT must align precisely with the first-order state causing it. The causal role of first-order states is discussed in further detail in this chapter.

condition are incompatible in a situation in which there is only one non-self-referential higher order representation (Block, 2011).

The strong (ambitious) version of HOT theory claims that phenomenal consciousness can be explained solely by the presence of a HOT, considering it sufficient for conscious experience. This approach faces issues when no first-order representation exists, as a HOT would suffice for conscious experience to occur, which requires further explanation about how this is possible. Moreover, this raises important questions regarding why our theory of consciousness should not be developed within a first-order representationalist framework if higher-order thoughts alone suffice to determine the nature of our conscious experience. If a thought (HOT) alone can make a mental state conscious, without its relation to first-order mental states, the theory ceases to be relational. This gives rise to the question why we do not defend a first-order thought theory. Following this line, one can apply the same kind of mechanism to perception and end up with a first-order representationalism, where the content of FOR makes the subject conscious.

On the other hand, the weak (moderate) version of HOT theory acknowledges that first-order representations (FORs) are necessary for consciousness, suggesting both HOT and FOR together are sufficient for conscious experience. This version either narrowly captures a technical sense of consciousness or considers FOR and HOT as jointly necessary for consciousness. Under such conditions, a theory proposing a constitutive mechanism linking first-order representations and higher-order thoughts might hold greater explanatory appeal. Indeed, self-representational theories – advocated by both critics of HOT theories, such as Neander and Block, and proponents within HOT theory, including Gennaro, van Gulick and Kriegel –

represent efforts precisely in this direction, emphasizing the essential relational structure between first-order and higher-order representations.

2.1.2 Further problems about misrepresentation

So far, I have talked about the negative outcomes of how HOT theory deals with misrepresentation. Problem 1 (P1) is that it gives HOT too much role in determining the nature of conscious experience, or in other words, the content of HOT overdetermines conscious experience. This is particularly evident when we examine the strong (ambitious) version of HOT theory. Strong HOT theory needs to explain the counter-intuitive outcome that one has a green experience both in hallucination and illusion, as this is the content of HOT, regardless of whether one sees a red object at the first-order level (case b) or has no sensation at all (case c). A related problem (P2) is defending HOT theory against a variant of first-order representationalism, as it turns out that a single mental state determines our conscious experience. This approach differs from the initial motivation of HOT theory, which elucidates consciousness by examining the relation between first-order and second-order mental states. It can be argued that we end up having a first-order theory of consciousness, where a thought (HOT) itself is responsible for consciousness. This could result in a theory suggesting that every thought is conscious, which does not align with the higher-order framework and poses significant issues. However, it is somewhat evident that not all thoughts are conscious. Sensory states, rather than intentional states, have given rise to discussions whether consciousness should be intrinsic to mental states (Rosenthal, 2005b, p. 119). A thought is only conscious if one is aware of being in that mental state and it is common for someone to have a thought, belief or desire, of which he is

not aware. It is also not the case that all higher-order thoughts are conscious. A mental state is only conscious if it is represented by an appropriate HOT. Typically, a HOT is not intransitively conscious. A higher-order thought becomes conscious only if a third-order thought represents it, which occurs infrequently in introspective consciousness (Rosenthal, 2005b, p. 127).

It is essential to establish a criterion to differentiate thoughts that induce consciousness from those that do not possess such capability. In this direction, Rosenthal postulates the extra condition that HOT should be subjectively non-inferential (Rosenthal, 2005, pp. 25, 57, 119). Bryne criticizes HOT Theories for introducing this extra condition, arguing that it seems arbitrary to assert that such a condition applies uniquely to higher-order thoughts: "… it is completely mysterious why a state's having (or lacking) a certain aetiology should be the extra ingredient that turns it into a state that there is something it's like to be in. And in any case, once we are allowed to appeal to aetiology, why not do it at the level of sensory states, leaving higher-order thoughts by the wayside" (Byrne, 1997, p. 123). Bryne addresses problem 2 by introducing an additional issue regarding the nature of HOT, questioning how its origin can influence the occurrence of consciousness (Problem 2b). We must substantiate why the etiology of HOT is significant, and if it is not, explain why such a distinction does not apply at the first-order level.

Another issue with the strong version of HOT theory, as Block highlights, is the absence of an object for a higher-order thought in cases of targetless misrepresentation, if we accept that conscious experience can occur without a first-order state. How can a HOT represent a non-existent first-order mental state? To address this, one must either revise the primary formulation of the theory or explain how a non-existent mental state can be represented. One option is to provide a

rationale for this counter-intuitive case; alternatively, one could revert to a weak version of HOT theory by either acknowledging the role of FOR or recognizing the presence of certain mental qualities essential for consciousness. In any case, further explanation for the relation between first-order representations and HOTs is necessary, clarifying the role of the first-order state without compromising the main traits of HOT theory (Problem 4). Additionally, a proper theory of mental qualities is needed, particularly addressing how we can make sense of unconscious mental qualities (Problem 5).

Problem 4 (P4) introduces new formulations of HOT theory and self-representationalism. These formulations aim to establish a constitutive and necessary relation between first-order representations and higher-order thought, thereby either preventing misrepresentation or presenting a plausible relational account in which both an existing FOR and HOT are integral. Problem 5 involves discussing the concepts of mental qualities, phenomenal character, and their connection to consciousness and what-it-is-likeness. This leads to the final problem (P6), which examines whether HOT theory and quality space theory can explain phenomenal consciousness in terms of what-it-is-likeness. The discussions on misrepresentation prompt questions about whether HOT theory only explains a form of technical or access consciousness, rather than phenomenal consciousness itself. Rosenthal's perspective on qualitative properties suggests that there can be mental qualities without consciousness, raising doubts about whether his views adequately explain phenomenal consciousness as the object of the hard problem.

At this stage, we can classify the related problems into six main categories: Problem 1: Higher-order thought overdetermines conscious experience, while the role of first-order representation is unsatisfactory and problematic. Problem 2: HOT

Theory risks collapsing into first-order representationalism because it explains consciousness through special thoughts that represent in a specific way. (This approach contradicts the theory's premise of explaining consciousness relationally and as *being represented*.) Problem 2b: Even if we demonstrate that the etiology of HOT is significant, how can we avoid the objection that the characteristics attributed to HOT are ad-hoc adjustments? Furthermore, how precisely does the etiology of HOT determine whether or not we have a conscious experience?[7] Problem 3: How can we explain state-consciousness of a non-existent first-order mental state? Problem 4: What is the exact nature of the relation between a first-order and second-order mental states? Problem 5: How can we explain an unconscious mental/phenomenal quality? Problem 6: HOT theory cannot adequately explain what-it-is-likeness.

2.1.3 Division of phenomenal labor

Neander's presentation of the problem through three scenarios demonstrate that division of phenomenal labor (DPL), the main strategy of HOT theories to explain consciousness and mentality independently, leads to various inconsistent positions. Beyond these listed problems, it is essential to assess whether division of phenomenal labor is a successful strategy and if HOT theory genuinely adheres to it. The main reason for adopting this approach is to provide different explanations for consciousness and mental qualities. Rosenthal explains consciousness through HOT

[7] If HOT theory gives a relational explanation, there should be two mental states between which a relation should be present. But the cases of hallucination show that only HOT is sufficient for consciousness to occur. How is this compatible with a relational approach? Can there be a relation between an existent and non-existent state? What is special about HOT to distinguish it from other thoughts, if only HOT is enough to constitute conscious experience? Why should we not move to a first-order thought approach rather than a higher-order approach?

theory and mental qualities with his quality space theory. A satisfactory theory of consciousness must distinguish consciousness from mentality; otherwise, it is uninformative and grounded on theoretical assumptions disguised as folk-psychological intuitions (Rosenthal, 2005, pp. 10, 22, 31).

Consciousness-based views and first-order theories align with the Cartesian paradigm, suggesting that consciousness is a component of all aspects of mental life. In the absence of consciousness, there is no experience, and consequently, no mental activity. Whenever experience occurs, consciousness is present, otherwise the concept of phenomenal experiences loses coherence. This approach takes consciousness to be an intrinsic feature of our mental states: It both fixes the reference of what mental is and also explains what mentality is. However, what fixes the reference of a notion is not necessarily the essential feature of that notion (Rosenthal, 2005, Ch. 5). Rosenthal argues that the initial premise of consciousness-based views is mistakenly portrayed as a folk-psychological intuition. On the contrary, he claims, it is based on Cartesian theoretical commitments. While our theory should be able to explain why we have such intuitions, it does not need to align with this idea that consciousness is intrinsic to mentality. For this reason, Rosenthal begins with the assertion that distinct explanations should be pursued. This methodological stance is used to support an ontological distinction between consciousness and mentality. Nevertheless, it remains methodically useful even if, ultimately, these phenomena prove interdependent. The concept of intrinsicality will be further explored in subsequent chapters.

The critics argue that HOT theory cannot pursue such a methodology consistently. If an overdetermining role is given to HOT, consciousness is explained only through HOT. In the typical misrepresentation case (scenario b), this would

mean the subject experiences green, aligning with the content of the HOT. Consequently, the theory collapses "qualitative character and subjectivity back together again" (Levine, 2001, p. 108). Conversely, if one maintains that the subject has reddish experience, an even greater challenge arises for HOT theory, as this would involve (a) collapsing two notions together (again), and (b) relying solely on single mental state for explanatory work. Both features risk reducing HOT theory to a first-order representationalism (Levine, 2001, p. 108).

To effectively use explanatory division of labor as a methodology, one should explain mental qualities and phenomenal consciousness independently. Neander argues that Lycan's explanation for mental qualities and consciousness conflict with each other.[8] Block also criticizes Rosenthal's explanation of phenomenal consciousness, arguing it fails to address what-it-is-likeness of experience, which is central to the hard problem and its associated mystery. According to Block, Rosenthal can only explain a form of higher-order awareness that does not illuminate phenomenal consciousness. Moreover, the possibility of misrepresentation highlights precisely this problematic feature of HOT theory. One either ends up explaining merely a form of higher-order awareness, insufficient to capture phenomenal consciousness, or encounters various difficulties when attempting to explain phenomenal consciousness through higher-order awareness – some of these difficulties were outlined in the previous section. The most important of these, to reiterate, is the occurrence of conscious experience as determined solely by the content of HOT in targetless misrepresentation. According to the division of phenomenal labor, consciousness is explained by the mechanism of higher-order awareness, even though no actual mental quality exists, leaving nothing to explain

[8] Original discussion involves qualia as mental qualities.

from first-order level. However, the critics argue that such a top-down explanation is insufficient because it dissolves the meaningful experiential distinction between successful and unsuccessful perceptions.[9] If consciousness were always a result of top-down representation, then perceptual and mental qualities would have no genuine impact on our conscious lives – an implication that appears untenable. Particularly problematic is the suggestion that conscious and unconscious pain would matter to us in the same way, despite the evident difference they make to our lived experience (Block, 2011).

Although Rosenthal offers an alternative explanation of mental qualities through his quality space theory, it remains ambiguous how these mental qualities can be considered phenomenal qualities if they lack consciousness. Conversely, if qualitative features as what-it-is-likeness could be elucidated through an account of mental qualities, a theory of higher-order awareness might not significantly contribute to the explanation of phenomenal consciousness. Thus, self-representationalism appears to be the only way to provide a relational and informative explanation. In this way, consciousness is explained both relationally – as a relation between two mental states which are constitutive parts of a single, complex mental state – and intrinsically, as an inherent qualitative feature of that unified mental state. Consequently, this approach dissolves the division of phenomenal labor and rather explains consciousness through two different aspects: as an intrinsic property and a relation between distinct mental states.

[9] In respect of how they seem and matter for us.

2.1.4 Overdetermination of HOT

We discussed that assigning HOT full responsibility can address the misrepresentation issue in conscious experience. We saw that this is problematic because it goes against the division of phenomenal labor and overlooks the role of first-order representation (FOR). Since HOT theory appears to ignore the role of first-order states in hallucination and illusion, this omission raises confusion regarding the nature of the role FORs play in veridical perception. As we do not want to disconnect our conscious perception from the world, we need to clarify the precise function that FOR serves within the higher-order framework.

Coleman thinks that HOT theory does not really "account for the consciousness of first-order states, since we are never strictly conscious of them" (Coleman, 2018, p. 40). He draws an analogy with a video installation in gallery, where we see the image of a painting through a screen on which the real painting's image is projected. In the case the painting is stolen, and the screen shows us the image of the painting that had been saved from the previous day, we would not experience any difference. This shows that the existence of the real painting does not make any difference on our experience whether it exists or not. This is analogous to saying that first-order mental state that is in contact with the real object in the world does not make any difference both in targetless misrepresentation and in veridical perception (Coleman, 2018, pp. 38-39).

The issue can be viewed as a conflict between intransitive and transitive consciousness, or between state and creature consciousness. Wilberg interprets HOT theory as a theory of state consciousness rather than creature consciousness: "it purports to articulate the conditions under which a mental state possesses the property of being conscious" (Wilberg, 2010, p. 624). Therefore, HOT theory

explains mental state *types* rather than *tokens*, and to claim that a subject S in a mental state M at time t, "one must say that the token state exists or occurs" (Wilberg, 2010, p. 624). The notion of state consciousness, seen as a property of mental state tokens and equated with phenomenal consciousness, is central in this framework as explicated by Rosenthal. In these terms, Rosenthal's explanation suggests that a mental state token becomes conscious in virtue of a HOT, through the subject's instantiation of a mental state type (HOT) at time t. However, Wilberg questions precisely which existing mental state token is conscious. Considering the conscious state as the object of the representational content of HOT does not solve the problem because the state itself may not exist. Consequently, characterizing the conscious state as a notional object conflicts with the transitivity principle, if we interpret TP as an explanation of mental state token, which must actually exist. The HOT theorist might respond by denying that one needs to be in any mental state, emphasizing instead that what matters is that the subject is conscious of oneself as being in a mental state, thus highlighting the divergence between appearance and reality and characterizing HOT theory primarily as a theory of appearance. Rosenthal clarifies this point by saying "being in a conscious state is not being in that state and being conscious of being in it, but simply being conscious of oneself as being in the state" (Rosenthal, 2004, p. 41). If we regard HOT theory as addressing state consciousness specifically, the implication is that in instances of targetless misrepresentation, a conscious mental state token simultaneously exists and does not exist – a scenario Wilberg finds clearly problematic.

HOT theorist might deny that HOT theory is, strictly speaking, an explanation of consciousness as a property of the first-order state. He might also assert that the theory does not genuinely provide a relational explanation between

FOR and HOT. Rather, the theorist can prioritize the transitivity principle, which underlies the overdetermination of conscious experience by HOTs. However, such a move pushes HOT theory toward a one-level account, transforming its characteristic two-layered relational structure into a single-level, one-directional representational relation determined entirely by a higher-order thought with a particular representational mode.[10] This outcome aligns more closely with first-order representational theories of consciousness, thereby weakening the distinctiveness of the higher-order thought approach.

2.1.5 Collapse into first-order theory

According to the transitivity principle and overdetermination of HOT, (a) a higher-order thought is sufficient to generate conscious experience, and (b) the content of the HOT fully determines the nature of this experience. This becomes particularly evident in cases involving misrepresentation. Targetless misrepresentation supports idea a, illustrating that the presence of HOT alone suffices for consciousness. Normal misrepresentation reinforces idea (b), indicating that the nature of the conscious experience is determined solely by the HOT's representational content. When our awareness of an object is shaped by how our higher-order thoughts represent our mental states, our conscious experience aligns strictly with the conceptual content within these HOTs. However, as previously discussed, in instances of weak misrepresentation – where a greenish first-order state is accompanied by a reddish higher-order thought – opting for either a green or red experience effectively pushes HOT theory toward a first-order representational formulation.

[10] The particular mode would be appropriate HOT: a non-inferential and seemingly immediate representation by a higher-order thought. It is similar to first-order representationalism, according to which one has a representation with a mental attitude.

Gois (2010) similarly argues that higher-order theories risk "collapsing into a first-order theory of consciousness by allowing HO states to introduce qualitative differences of their own" (p. 155). Accordingly, the two-layered structure is effectively abandoned. If higher-order thoughts alone determine the qualitative features of our conscious experiences, rather than first-order states, then the division of phenomenal labor and the two-layered structure are rejected. Consequently, both phenomenal qualities and consciousness become explicable solely in terms of HOTs, resulting in a form of first-order representationalist theory, in which representing things as having certain properties becomes sufficient for consciousness.

It is important to note that first-order representational theories and higher-order representational theories differ in their approach to the direction of representation. While first-order representational theories assign the explanatory role to representing, higher-order representational theories focus on state's being represented (Kriegel, 2009). Rosenthal's use of the transitivity principle appears to support such a first-order representational view, because it explains consciousness through a HOT's representing oneself to be in a mental state M. Once again, this reveals a tension between the division of phenomenal labor and the transitivity principle. HOT theory appears to be incapable of providing an explanation that aligns with both concepts.

2.1.6 Non-existent first-order state

Another important inconsistency resulting from the overdetermining role of HOT is the assumed possibility that higher-order thought can represent a non-existent first-order mental state. Block identifies this as an inconsistency between sufficient and necessary conditions of HOT theory. Such criticisms are based on an understanding

of HOT theory in the following way: "A mental state is conscious if and only if the state is the object of a certain kind of representation arrived at non-inferentially" (Block, 2011, p. 421). We can find similar descriptions of the theory in the literature: "Higher-order representational theorists argue that a mental state M is phenomenally conscious if it is appropriately represented by a higher-order state M*" (Picciuto, 2017, p. 556).

Accordingly, radical misrepresentation involves a HOT, and conscious experience should occur only if the HOT accurately represents its object (first-order state) in an appropriate way. However, in targetless misrepresentation, there is no first-order state, thus no object to be represented by a HOT. Considering that in HOT theory the higher-order thought is not self-representational (it does not represent itself), this scenario implies a non-existent first-order state being represented by a HOT to be conscious. There is a conflict between the necessary condition for consciousness that there should be a first-order state alongside the higher-order thought, and the sufficient condition that an appropriate HOT is enough for consciousness to occur. Possibility of targetless misrepresentation shows that both conditions cannot be satisfied at the same time and for this reason "higher-order approach to consciousness is defunct" (Block, 2011, p. 419).

Picciuto considers the problem to be dependent on one's approach to division of phenomenal labor (DPL). Picciuto argues that DPL is an essential feature of the HOT theory along with transitivity principle (TP) and *extrinsicality*. According to him, one can "hold the division of phenomenal labor with varying degrees of strength" (Picciuto, 2015, p. 561). A weak endorsement of division of phenomenal labor does not commit HOT theories to the existence of first-order state as a necessary condition for conscious experience, which Picciuto presents as the

"existence condition (EC): A subject S undergoes a phenomenally conscious state if and only if S undergoes an actual existing first-order state that is represented by a higher-order state" (Picciuto, 2017, p. 560). To clarify further, the central aim of DPL is twofold: it elucidates qualitative features through first-order representations and explains the consciousness through higher-order thoughts. Consequently, HOT theorists such as Rosenthal and Weisberg, who view HOTs as sufficient for consciousness, can coherently commit themselves to a weak version of DPL. Under this weaker interpretation, consciousness is determined by HOTs, whereas qualitative features are determined "by the first-order state as it is represented by the higher-order state, even though that first-order state could be an intentional inexistent" (Picciuto, 2017, p. 561).

Let’s revisit the conflict between the transitivity principle (TP) and the division of phenomenal labor (DPL). TP dictates that a mental state is conscious whenever the subject is aware of being in that state. For Rosenthal, TP is the essential starting point for HOT theory and offers the most plausible account of consciousness. It is based on the commonsense idea that a mental state is not conscious if the subject is in no way aware of being in it. However, the status of TP in HOT framework is arguable, as Gennaro takes it as a conceptual truth and Lycan presents it "as a stipulative definition that corresponds to at least one commonsense notion of consciousness" (Picciuto, 2017, p. 557). Transitivity principle is based on the transitive consciousness and explains phenomenal consciousness through the awareness-of relation between the subject and his mental states. On the other hand, DPL aims to give an account of intransitive state consciousness through a representational relation between two mental states (FOR and HOT) independent of an account of mental qualities of first-order states. Moreover, one might characterize

TP as focusing primarily on creature consciousness, thus viewing phenomenal consciousness as a property of subjects, whereas DPL emphasizes state consciousness, thereby viewing it as a property intrinsic to mental states themselves. The division of phenomenal labor is notably problematic in cases of misrepresentation, where TP seems initially to offer a solution. However, TP introduces further complications because, in certain discussions, it appears sufficient to account for phenomenal consciousness independently, without necessitating the existence of FOR. Thus, one might reasonably interpret the problem as a conflict between these two central principles.

Critics such as Block argue that a higher-order thought (HOT) cannot overdetermine conscious experience in the absence of a corresponding first-order state. Accordingly, it is unacceptable to assert that a non-existent pain matters in the same way as the existent pain (Block, 2011, p. 427). This discussion is based on the application of misrepresentation on the empirical case of dental pain as discussed both by Rosenthal and his opponents: One can consciously experience dental pain triggered merely by fear and the sound of drilling before any physiological stimulation occurs. HOT theorists view this as a positive instance of targetless misrepresentation, demonstrating that radical misrepresentation is not just a theoretical possibility but also empirically plausible, and that HOT theory can successfully anticipate and explain such cases. Block interprets this claim as suggesting that non-existent conscious pain matters as much as actual conscious pain if one accepts the HOT theory's principle of overdetermination, a stance he finds implausible.

The absurdity of this position becomes clearer when considering a modified version of Neander's three-scenario example: Suppose in all these three cases, one

has a HOT that one is in agonizing pain. However, while (i) one has a proper first-order state representing pain in the first scenario, (ii) a first-order state representing the taste of chocolate exists in the second scenario; and (iii) no first-order state exists at all in the third scenario. Block questions how one could rationally claim that the conscious experience is identical in these distinctly different conditions. He argues it is absurd to equate the significance of a first-order representation (FOR) involving actual agonizing pain (i) with a FOR involving the taste of chocolate (ii) or with the absence of any FOR altogether (iii). He concludes that HOT theory can only survive in a modest form, focusing on higher-order consciousness rather than fully accounting for the what-it-is-likeness of experiences. For Block, it does not make sense to talk of consciousness with the aim of explaining what-it-is-likeness without considering first-order qualitative features as part of consciousness.[11]

Weisberg (2011) formulates the problem of misrepresentation with respect to non-existent lower-order state through the following misrepresentation argument:

> 1. If there's something it's like for the subject, the subject is in a conscious state.
> 2. In radical misrepresentation, there is something it's like for the subject.
> 3. So, in radical misrepresentation, the subject is in a conscious state.
> 4. The conscious state is either the first-order state or the higher-order state.
> 5. The first-order state does not exist in radical misrepresentation.
> 6. The subject is not aware of the higher-order state in radical misrepresentation.
> 7. So, either the subject is in a conscious state that does not exist, or the subject is in a conscious state that she is not aware of being in.
> 8. A subject can't be in a conscious state that does not exist.
> 9. So, the subject is in a conscious state that she is not aware of being in. (p. 417)

Premise 1 follows from the Nagelian definition, and premise 2 derives from Rosenthal's view that consciousness is a matter of mental appearance, as well as from the possibility of conscious experience in radical misrepresentation. Premise 3

[11] Compare with higher-order responses: 1) Consciousness is a matter of mental appearance. 2) What-it-is-likeness is not a common currency.

logically follows from P1 and P2. Weisberg identifies premises 4 through 6 as representing radical misrepresentation within HOT theory. P7 emerges naturally: If the first-order state is conscious, it must be either non-existent (the first part of P7), or the subject must not be aware of being in that conscious state (the second part of P7). One might propose conscious higher-order thoughts (HOTs) as a third disjunct within premise 7, which would significantly alter the conclusion. However, attributing consciousness directly to HOTs conflicts with the core principles of HOT theory. HOTs, as unconscious states, are designed to render first-order states conscious (or alternatively, make the subject conscious of their first-order states). This feature is one of the theory's key advantages, providing a naturalistic account of consciousness through unconscious HOTs. Without this constraint, HOT theory would risk infinite regress, as the consciousness of a HOT would require a third-order representation, leading to an infinite sequence.

Weisberg suggests that opponents of the HOT theory consider premise 8 self-evident, thereby concluding premise 9. What is important is that higher-order thought becomes irrelevant to explain consciousness of mental states (of both FOR and HOT) in targetless misrepresentation. If FOR cannot be conscious due to its non-existence, and the HOT itself cannot become conscious, then the only remaining possibility is that the subject is in a conscious state without being aware of being in or not in a conscious state at all. Consequently, the subject finds himself in a conscious state, for which a higher-order thought is not responsible. This means that the transitivity principle is not at work, as the subject remains unaware of his conscious state.

HOT theorist can address this limitation by emphasizing (a) that he situates consciousness at the level of subject's consciousness and (b) the relation between HOT and FOR. The first move would be based on TP without the involvement of

DPL. It would explain subject's consciousness with minimal talk of states. It is essential to demonstrate how this is feasible, as it contradicts a significant aspect of the theory that aims to explain state consciousness. The second option would need to justify how a relation can obtain when one part of the relation (FOR) does not exist. This raises similar issues regarding the non-existence of first-order representation.

Another option for HOT theorists is to deny P8 and argue that a subject can indeed be in a conscious state that does not exist. Rosenthal and Weisberg explicitly argue in this direction. Although they consider this move both natural and straightforward, it requires proper justification. First, it is important to clarify the role of FOR and explain the nature of HOT's object in veridical perception and hallucination. Does FOR play a role in (conscious and unconscious) veridical perception, only to be substituted by another mechanism during hallucination, or does it lack an essential role altogether in both cases? Furthermore, is the first-order perception always the object of HOT, or can there be instances where they do not converge? Second, what precisely does it mean for a FOR not to exist; in other words, how should we interpret the intentional object of a higher-order representation when that object is absent? Addressing this requires further discussion about whether the object of higher-order representation is first-order state. To adequately handle this issue, a proper account of intentional inexistence is necessary.

2.1.7 What-it-is-likeness

Farrell follows Block's distinction between the *modest* HOT theory that explains what-it-is-likeness (WIIL) only in a technical higher-order sense and the *ambitious* HOT theory that attempts to explain what-it-is-likeness in its traditional experiential sense as phenomenal consciousness. He argues that HOT theory oscillates between a

non-standard HO reading of WIIL (similar to Block's modest HOT theorist) while dealing with misrepresentation problem, and a standard reading of WIIL (similar to Block's ambitious HOT theorist) while utilizing the transitivity principle in awareness argument.

Farrell (2017) argues that for the HOT theorist to be considered ambitious – rather than modest and concerned only with a technical sense of consciousness – he must explain WIIL in accordance with the Nagelian definition: "M is conscious iff there is something it is like for S to be in M" (p. 2744). To do so, the theorist needs to deny the misrepresentation argument, according to which the subject experiencing radical misrepresentation is purportedly in a mental state that does not exist. On the other hand, HOT theorist must affirm the awareness argument, which shows that mental states are conscious through their subjects' being aware of them, following the transitivity principle "M is conscious only if S is aware of M" (Farrell, 2017, p. 2744). However, this position can only be maintained by equivocating on the notion of WIIL within these two arguments. To demonstrate the different usages of WIIL, Farrell introduces two distinct notions:

> SOMETHING: There is something it is like for S to be in M.
> WHAT: There is an occurrence of what-it-is-likeness associated with M. (Farrell, 2017, p. 2745)

Farrell presents the Misrepresentation Argument (Argument M2) as follows:

> (M1) M is conscious & S is not in M
> (M2) If M is conscious, there is something it is like for S to be in M
> (M3) If there is something it is like for S to be in M, there is an occurrence of what-it-is-like-ness associated with M
> (M4) If there is an occurrence of what-it-is-like-ness associated with M, S is in M
> (M5) If there is something it is like for S to be in M, S is in M. (From (M3, M4).)
> (M6) S is in M. (From (M1), (M2), (M5).)
> (M7) S is not in M. (From (M1).)
> (M8) S is in M & S is not in M. (From (M6), (M7).) (Farrell, 2017, p. 2746)

M2 is Nagelian definition of WIIL. Farrell considers M3 to follow naturally from SOMETHING and M4 to follow from WHAT. M5 logically follows from M3 and M4. Consequently, we arrive at the contradictory conclusion that S both is and is not in M. Therefore, the initial assumption M1 must be wrong. It is thus impossible for the subject to be in a conscious mental state that does not exist as HOT theory proposes.

Farrell argues that HOT theorist must reject M5 to argue against the misrepresentation argument, which necessitates rejecting either M3 or M4. Farrell maintains that HOT theory accepts M3 because it endorses an *occurrent* reading of SOMETHING, which implies the occurrence of WIIL associated with M. He points out that the standard reading of SOMETHING is *non-occurrent*, meaning HOT theory departs from the conventional reading by following a non-standard reading. The non-occurrent reading of SOMETHING does not entail that if there is something it is like for S to be in M, then there must be an actual occurrence of WIIL associated with M at time t. This distinction can be best illustrated through the thought experiment Mary the scientist. According to non-occurrent reading, if Mary were to leave her room at time t, she would actually be seeing a red object. Therefore, even while she remains in her black-and-white room (prior to actually leaving at time t), there is something it is like for Mary to see red. Thus, the non-occurrent reading suggests that there can be something it is like to see red, even in the absence of an occurrent experience of redness.

For this reason, HOT theory needs to deny M4. HOT theory adopts a *loose reading* of WHAT, which does not imply that S is in M. According to this interpretation, an occurrence of what-it-is-likeness does not require the subject to be in that mental state. This is again a non-standard use of WHAT in comparison to the

tight reading, in which the occurrence of what-is-like-to-be-in-pain requires the subject to be in relevant pain state. For HOT theorist, Farrell argues, this is not necessary: If one represents oneself as having a migraine, there is an occurrence of WIIL associated with migraine, even if S is not in that mental state. "This representing can occur without the migraine occurring" (Farrell, 2017, p. 2750). As a higher-order thought is sufficient for conscious experience, consciousness can occur independently of one's actual presence in the first-order mental state. Thus, HOT theory rejects M5 by denying M4.

This presents the first significant problem for the HOT theorist: The misrepresentation argument can only be countered by denying M4 and adopting a loose interpretation of WHAT. However, this interpretation makes it possible to talk about an occurrence of what-it-is-likeness without the subject being in the relevant first-order state. Philosophers like Ned Block regard this as reflecting merely a technical or higher-order sense of consciousness, which only a modest HOT theorist can pursue. Such a reading collapses the distinction between real and illusory what-it-is-likeness, undermining the very explanation of consciousness that HOT theorists aim to provide. Consequently, HOT theorist denies misrepresentation argument at the expense of giving up his ambitions.

Farrell (2017) continues his general argument by asserting that this reading is in conflict with the higher-order reading of WHAT employed in awareness argument (specifically in premise M3), which he presents as follows:

> (A1) If M is conscious, then there is something it is like for S to be in M
> (A2) If there is something it is like for S to be in M, then S is aware of M
> (A3) If M is conscious, then S is aware of M. (From (A1), (A2).) (p. 2751)

To affirm the awareness argument, the HOT theorist needs to justify premise A2. Farrell argues that the HOT theorist cannot achieve this by relying solely on the

transitivity principle, as doing so would beg the question. Therefore, the HOT theorist must employ a *self-intimating* reading of SOMETHING, which entails that S is aware of M. Farrell argues against employing a non-standard reading of SOMETHING.

The resulting problem is that the HO theorist adopts a non-HO reading of SOMETHING in A1 and M2, while applying a HO Reading of SOMETHING in A2. Premise A1 follows from Nagelian definition, which utilizes a non-HO interpretation of SOMETHING. According to this interpretation, it is not necessary that WIIL requires a higher-order state for a subject to become aware of his mental state M. Farrell argues that Nagel's primary concern is not the awareness of the subject per se, but rather the refutation of physicalism. Nagel focuses specifically on an experiential property that cannot be adequately captured through the objective methodologies of physics. On the other hand, premise A2 relies explicitly on a HO reading of SOMETHING; otherwise, the affirmation of the awareness argument would be uninformative and circular. Consequently, the HOT theorist applies different readings of SOMETHING across different premises. In conclusion, even if HOT theory is not defunct – as Block argues – it nonetheless appears to shift its explanatory focus and thereby fails to maintain an overall consistent position regarding both the misrepresentation argument and awareness argument.

2.2 Dealing with the problem of misrepresentation

In the first section, I detailed the problem of misrepresentation as a critical challenge for higher-order representational theories, exploring its primary features and related implications within this theoretical framework. In this section, I will address

potential responses available within the higher-order paradigm to counter these criticisms.

2.2.1 Possibility of misrepresentation

Rosenthal argues that misrepresentation is both (a) a theoretical possibility and (b) an actual occurrence (Rosenthal, 2005, p. 166). The potential for misrepresentation, as a theoretical implication of HOT theory, should not be viewed negatively; rather, it constitutes an advantage. Misrepresentation frequently manifests in various forms, and HOT theory provides a more robust explanation of these occurrences compared to alternative frameworks.

Most philosophers criticize HOT theory due to its susceptibility to targetless misrepresentation. In response, alternative formulations have been proposed to address and prevent these occurrences. However, HOT theory is not the only theory that allows for misrepresentation: "any theory on which a mental state's being conscious consists in one's being transitively conscious of that state will do so. As long as a conscious state is distinct from one's being transitively conscious of it, the content of that transitive consciousness may misrepresent the state" (Rosenthal, 2002, p. 415). Therefore, the transitivity principle inherently permits misrepresentation. As consciousness of a mental state is determined by one's awareness of that mental state (independently from the mental qualities of that state), we will be conscious of whatever content we represent ourselves as being in. For this reason, it is possible that the way we represent ourselves might fail to correspond to the content of our perception, and it could even be the case that the relevant first-order state is entirely absent. For Rosenthal, the distinction between misrepresentation and targetless misrepresentation is arbitrary.

It is unclear precisely which theories allow for misrepresentation. The transitivity principle is the central component common to all forms of higher-order theories and self-representational theories. The central premise of these theories is that intransitive state consciousness can be explained in terms of transitive state consciousness; in other words, the property of a mental state being conscious is explicated through the relation of being represented by another mental state. Although self-representational theories also utilize TP, the majority of their proponents argue that their theories effectively avoid the issue of misrepresentation. Thus, merely employing transitive consciousness and the transitivity principle does not automatically entail the possibility of misrepresentation. The critical issue here is not the acceptance of TP itself, but rather the understanding of its underlying motivation: Consciousness should not be considered an intrinsic quality of a conscious mental state. Rather, it should be explained through a representational awareness-of relation established by an extrinsic mental state. Distinctness and extrinsicality are the key criteria here. Self-representational theories attempt to implement both transitivity principle and an intrinsic understanding of state consciousness. As they use an intrinsic concept of consciousness, their proponents argue that they prevent misrepresentation. However, insofar as they still incorporate the transitivity principle, their theories remain susceptible to misrepresentation. Demonstrating that consciousness is not intrinsic to mental states thus constitutes the fundamental disagreement between HOT theory and its opponents in numerous debates, from the problem of misrepresentation to the hard problem of consciousness.

As Rosenthal contends, most of the objections based on the problem of misrepresentation regard no form of higher-order theory as satisfactory "unless it

rules out misrepresentation on its own, without any added provision" (Rosenthal, 2018, p. 53). There is no independent and non-stipulative reason for insisting that we must eliminate misrepresentation entirely. Thus, either the motivation to preclude misrepresentation is justified, or it is being used strategically to critique other features of HOT theory, independently of misrepresentation itself. For most critics, defending an intrinsic conception of consciousness constitutes the primary motive behind their use of the possibility of misrepresentation to challenge HOT theory. For HOT theorists, this same motivation underlies criticisms regarding overdetermination of HOT, the lack of role for first-order state on consciousness during veridical perception, and the absence of first-order state in targetless misrepresentation.

2.2.2 Intrinsicality

Intrinsicality is the idea that consciousness is an intrinsic property of mental states and indistinguishable from mental qualities. HOT theorists accuse their opponents – generally the philosophers who think there is a hard problem of consciousness – of using an intrinsic notion of consciousness without giving an independent justification. Ned Block explicitly endorses this intrinsic reading, stating that "sensory qualities have a kind of consciousness in themselves" (Block, 2011), referring here to phenomenal consciousness or what-it-is-likeness that is intrinsic to the mental state. On the other hand, HOT theorists, following the division of phenomenal labor, argue that mental qualities can exist without the presence of phenomenal consciousness. Thus, there is no what-it-is-likeness associated with these states. A mental state is conscious only if the subject is aware of himself as being in that state (through appropriate HOT). If this is not the case, then there would be nothing it is like to experience those unconscious first-order states.

For Block, higher-order approach presents the conditions "in which an actual state is a conscious state" (Block, 2011, p. 427), but does not provide any substantive account on what consciousness or WIIL is. We can see this better when we want an explanation from HOT theorist about what would happen in the targetless representation of the non-existent/empty mental state. According to Block, the fact that a conscious but fake/non-existent pain would still matter experientially in the same way supports his criticism of HOT theory – namely, that the theory either explains a technical and uninteresting sense of WIIL or is inconsistent.

Weisberg argues that misrepresentation is a problem insofar as one relies on an intrinsic characterization of consciousness. Accordingly, phenomenal consciousness is intrinsic to the qualitative properties (as being closely associated with the theoretically problematic notion of qualia), which are thought to be "independent from causal, functional, or intentional features of the mind" (Weisberg, 2011, p. 410; Block, 1995, p. 234). In contrast, an extrinsic characterization is open to a functional explanation of consciousness within a naturalized representationalist theory. This characterization acknowledges both the special role consciousness plays in our cognition (and mental systems more broadly), and the folk-psychological understanding of consciousness as intrinsic, thereby also explaining why we tend to conceive it intrinsically. Weisberg elsewhere distinguishes between a modest and zealous reading of consciousness, paralleling the extrinsic-intrinsic distinction. According to Weisberg, Block pursues a zealous reading of phenomenal consciousness that is intrinsic and mysterious, while one can pursue a neutral and moderate reading of consciousness that just means experience (Weisberg, 2011b). The important points for HOT theorist are: (a) The intrinsic characterization of consciousness is heavily theoretical and still awaits independent justification; and (b)

our intuitions do not determine the reality. Even if we do not want to explain them away, they can be revised if they conflict with a successful theory.

If we examine Weisberg's formulation of misrepresentation argument (see argument 1 in this chapter), we can observe that the HOT theorist can reject premise 8 (a subject can't be in conscious state that doesn't exist). According to Weisberg, this rejection can be grounded in adapting an extrinsic characterization (Weisberg, 2011, p. 419). Transitivity principle employs an "intentional construction of 'aware of oneself as'…", which does not require the intentional object to exist (Weisberg, 2011, p. 418). For instance, I can be aware of myself as a twenty-year-old Cuban pianist, even if I am not. Conceiving myself as being in a state with particular properties does not thereby bring those states and properties into existence.[12] In this sense, the transitivity principle "introduces an intensional context", and consequently, the objects of HOTs are intentional objects of representation, which do not exist (Weisberg, 2011, p. 419). This issue arises only if one bases their account on an intrinsic notion of consciousness.

I believe the distinction between extrinsic and intrinsic conceptions of consciousness play an important role in the discussions concerning the problem of misrepresentation. However, it remains questionable whether Weisberg's distinction provides a definitive solution. Firstly, adequately characterizing the object of HOT requires additional theorizing. One cannot simply dismiss the idea of intentional inexistence as absurd without further argument. It is also important to consider the nature of objects in higher-order thought during radical misrepresentation, as well as the complex implications this has for successful perception. Secondly, Block situates

[12] We cannot say the same about the subject of the experience. I must exist, but the states I am aware of myself as being in and the properties of my conscious perception do not have to be present or exist in the way I am conscious of them. I am grateful to Lucas Thorpe for raising this issue.

consciousness within a framework that views it as inseparable from mental qualities. He adopts a functionalist perspective and supports a view of consciousness that is both functional and intrinsic. Moreover, he argues that most functionalist and representational theories of consciousness, which take the phenomenal dimension seriously, endorse this intrinsic view. Thirdly, the conflict between necessary and sufficient conditions of HOT theory is not totally dependent on the intrinsic characterization. Fourthly, the role of first-order state in veridical perception and the role of HOT on our behavior remain matters of debate. In this direction, fixing the reference of consciousness with higher-order theories' extrinsic concept might not be enough to deny the misrepresentation problem.

Demonstrating that defining consciousness with an intrinsic concept does not contribute significantly to our theoretical understanding may provide a rationale for favoring an extrinsic characterization. Rosenthal argues that we fix the reference for mental qualities in the same way, independent of whether they are conscious or not. On the other hand, we can fix the reference of consciousness by adapting a suitable extrinsic characterization. This method enables us to determine mental qualities and consciousness independently. In contrast, the intrinsic characterization implies that consciousness and mentality must be defined simultaneously. The idea is that it does not make sense of talking about qualitative features without their being conscious. However, qualitative features are qualitative not in virtue of being conscious, but rather due to their inherent qualitative nature – that is, their distinctiveness from the physical properties of the objects they represent. They are differentiated representationally with the way they are able to make discriminations in our perception, paralleling the perceptible qualities present in external objects. These distinctions remain regardless of whether a qualitative property is conscious or not.

The intuition that qualitative properties and consciousness are inseparable arises from the fact we can only refer to qualitative properties when they are conscious. Nonetheless, the fact that we possess first-person access only when these mental states are conscious does not entail that these states are not qualitative when they are unconscious.

Common sense provides significant support for critics since it is intuitively difficult to conceive of a subject as being in a conscious state that does not exist (Weisberg, 2011, p. 419). However, as Rosenthal argues, if one shows that a given theory provides the best explanation of the explicandum, then contrary intuitions alone are not sufficient to reject the theory. It is not unseen that a successful theory leads to a revision or denial of a commonsense intuition (Weisberg, 2011, p. 421). Also, "common sense allows that we make errors about the conscious states we are in" and there is no "principled way of ruling out more radical mismatches" (Weisberg, 2011, pp. 419-420). Thus, there is no strong reason to reject an intensional reading of transitivity principle. One might conclude that criticisms based on misrepresentation gain appeal primarily insofar as one uses an intrinsic conception of consciousness, while HOT theory remains strong in so far as it adopts the transitivity principle. For this reason, the justification for these opposing positions constitutes an important criterion in deciding between theories and our position toward misrepresentation.

It is important to acknowledge that the intrinsic concept of consciousness aligns with several common theoretical commitments concerning our mental states: (a) Our mental states are infallible; (b) our mental states are transparent to us; we have direct access to their contents and therefore possess better knowledge of our mental states than others do; (c) all mental states that have qualitative character also

have the property of consciousness; and (d) in matters of consciousness, there is no distinction between appearance and reality as there is no mental reality beyond our being conscious of them.

2.2.3 No-consciousness in misrepresentation

We have seen that higher-order (HO) theories explicitly commit themselves to some form of overdetermination of HOT. Rosenthal writes: "What it's like for one on these theories is determined by the way the higher-order awareness represents the first-order state. Consciousness is a matter of mental appearance, that is, of how our mental lives appear to us, and on those theories that mental appearance is due solely to the higher-order awareness" (Rosenthal & Weisberg, 2008, Sec. 5). There are two key claims here: (1) The content of HOT determines the nature of conscious experience, and (2) consciousness is a matter of appearance. Clearly, one should understand (1) with the related claim (1a) that the occurrence of consciousness is dependent on the presence of an appropriate HOT. Similarly, (2) aligns well with the related claim (2a), indicating a divergence between appearance and reality. We have already discussed classical problems related to the overdetermination of HOT and Wilberg's position, which locates the core of the problem in an insufficient explanation of state consciousness as a mental state token. He does not suggest abandoning HOT theory but rather argues that it is better to conclude that there is no conscious experience in targetless misrepresentation. Rocco Gennaro follows a similar line of thought, emphasizing that the alternative – that no conscious experience occurs in misrepresentation – is dismissed prematurely, even though it remains a viable option. Both Wilberg and Gennaro attempt to show that Levine is

wrong by saying this option is ad hoc and not suitable to a HO framework (Levine, 2001, p.108).

There are two views on the difference between moderate and radical misrepresentation. Whereas Rosenthal argues that the distinction is arbitrary (Rosenthal, 2004, p. 32), Wilberg finds it helpful to treat them independently (Wilberg, 2010, p. 620). Wilberg defines moderate misrepresentation in HO terms as (MR1) "HOTs which represent, roughly accurately, the individuating features of some existing mental state", and radical misrepresentation (MR2) as “HOTs which do not represent, roughly accurately, the individuating features of any existing mental state" (Wilberg, 2010, p. 621). The following discussions will focus particularly on targetless misrepresentation.

Wilberg considers a scenario – originally suggested by Rosenthal – involving a dental patient who experiences dental pain following the sound of drilling and feelings of fear, despite the absence of any actual stimulation of the teeth (and thus no corresponding first-order sensation of pain). Wilberg treats this scenario as a genuine case of targetless misrepresentation. He argues that HOT theorist has four options to deal with the targetless misrepresentation: (i) When there is no existing first-order state to be conscious, HOT itself might be regarded as conscious; (ii) one can consider other stimulating states as causing the HOT to represent appropriately – in dental pain scenario these would be the sound of drilling and/or fear to cause a HOT to represent the subject to be aware of himself as having a dental pain; (iii) the subject's consciousness could be determined entirely by the representational content of the HOT itself; and (iv) no conscious experience occurs at all. According to Wilberg, options (i) and (ii) are not suitable for a HOT response: Option (i) leads to an infinite regress, as it would require a third-order state to make the HOT conscious;

it is based on an intrinsic idea of consciousness. Regarding option (ii), if we follow Wilberg's description of a HO reading of targetless misrepresentation, the HOT would fail to represent roughly accurately the individuating features of M, because "the HOT I am in pain is not even roughly accurate as a representation of my sensation of drilling, or my fear" (Wilberg, 2010, p. 623). Option (iii) is not satisfactory due to the problems we have discussed as part of overdetermination of HOT. Therefore, Wilberg argues that option (iv) is the best available option, even though it suggests some revisions to the classical HOT Theory.

Wilberg's justification is based on a modified version of MR2 as a main principle for a HO explanation: “If a mental state token (S,M,T) is conscious in virtue of a HOT, the HOT must represent, roughly accurately, the individuating features of M” (Wilberg, 2010, p. 627). Accordingly, there is no existing mental state token that is roughly accurately represented by a HOT in terms of its individuating features. This means there is no mental state token that is conscious and thus no conscious mental state. Wilberg responds to the possible criticism that this position might seem counterintuitive by urging a distinction between counterintuitiveness of the possibility of targetless misrepresentation from the counterintuitiveness of there being no conscious experience in such cases. One can see that any apparent counterintuitiveness arises primarily from the very notion of targetless misrepresentation itself, rather than from Wilberg’s own suggested solution. If one finds the possibility of radical misrepresentation implausible, either HOT theory does not need to defend itself against the misrepresentation problem, or there is something fundamentally wrong within HOT theory, since it allows such possibilities. Once we accept that misrepresentation is indeed possible and that actual cases exist, it is not a big step to argue that no conscious experience would occur under these conditions.

We should also be careful not to infer that the subject would lose all of its consciousness in targetless misrepresentation. It is possible to have multiple simultaneous HOTs, and in cases of targetless misrepresentation, only one HOT might be targetless, resulting specifically in the absence of consciousness associated with that state. Yet, if the subject has another HOT appropriately representing the individuating features of a different mental state (M2), then the subject will be conscious as represented by the successful HOT (Wilberg, 2010, p. 628). The counter-intuitiveness arises solely from the idea that the subject would not be in any conscious mental state, which would only be possible if the subject had only radically false HOTs at time t, and this is mostly not the case.

Wilberg suggests a modified version of HO theory that can successfully deal with targetless misrepresentation at the cost of rejecting two of Rosenthal's main principles: (1) Instead of saying that consciousness is a matter of appearance, he argues that there is a distinction between reality and appearance in consciousness; and consciousness is a matter of veridical appearance; (2) instead of saying that a subjectively non-inferential HOT is enough to satisfy the requirement of an appropriate representation of HOT, he argues that HOTs need to be veridical. Wilberg emphasizes a rough correspondence between appearance and reality and highlights the importance of the etiology of HOT in determining successful conscious experience. Here, a distinction between the constitutive and merely causal is illuminative. Accordingly, HOTs could have causal roles to produce conscious states in hallucination or various placebo effects. However, it is inaccurate to state that we have conscious experiences due to the misrepresenting higher-order thought (HOT) in the constitutive sense. One would only have a conscious experience in

virtue of a HOT in constitutive sense if that HOT roughly accurately represents the individuating features of M.

Wilberg's approach qualifies as a higher-order theory because it follows the general HO principle that a conscious state is conscious insofar as it is represented by a HOT. He expands it and concludes that "a conscious state is conscious only in those respects in which it is (roughly) accurately represented" (Wilberg, 2010, p. 634). The usage of this principle in his treatment of moderate misrepresentation is particularly interesting. Consider the change blindness experiment, where a subject perceives only a green parrot while his visual sensation alternates rapidly between a green and red parrot – rendering the subject unable to detect the difference in stimuli (Grimes, 1996). Specifically, examine the situation at time t1, when there is a visual sensation of a red parrot and a conscious experience of a green parrot. Classical HOT response to this case is to say that the subject is conscious in relation to the content of his HOT – he represents himself as seeing a green parrot. However, Wilberg opposes, according to such a picture, one would be conscious both of a red and green parrot. The first-order visual state would be conscious, because the subject's HOT roughly accurately represents the individuating features of M, in terms of its parrotness. At the same time, one would roughly accurately represent the inexistent mental state of his HOT in terms of its redness. Both parrotness and redness might be considered as the individuating features of M as being roughly accurately represented. However, we could suggest a better solution: "the visual experience of a red parrot is conscious in virtue of the HOT that one is having a visual experience of a green parrot, it is not conscious in respect of the redness of the parrot" (Wilberg, 2010, p. 634).

Although Wilberg's solution to the targetless misrepresentation is promising, his treatment of moderate misrepresentation is confusing. It is not clear whether one has an experience of a parrot with indeterminate color or a green color. If both are possible, then one should explain the difference in HO terms. Also, as change blindness experiments show that the subjects occasionally have the experience of a green parrot, Wilberg needs to explain how the greenness is attained. His view only explains how the property of parrotness is acquainted.

Gennaro also tries to show that no-consciousness option is a viable scenario in the classical formulation of Levine's misrepresentation argument. Levine had considered four options for a possible case of misrepresentation involving a visual sensation of red diskette case accompanied by a HOT representing it as green. Option 1 is a conscious greenish experience, option 2 is a reddish conscious experience, option 3 is no resulting conscious experience and option 4 ensures "correct representation by pinning the content of the higher-order state directly to the first-order state" (Levine, 2001, p. 108). Gennaro argues that both option 3 and option 4 are plausible routes available to HOT theorists. Gennaro, like Wilberg, emphasizes that Option 3 does not imply total absence of consciousness, which would mistakenly conflate creature consciousness with state consciousness (Gennaro, 2004, p. 58). Rather, it just means that the subject would not have a greenish or reddish conscious experience. For him, this is not an ad hoc move but a very legitimate possibility arising precisely in cases where "there isn't a match between the contents of the lower-order and higher-order states" (Gennaro, 2004, p. 58). Radical misrepresentation scenarios presented through thought experiments often lack detailed descriptions of precisely how a HOT misrepresents. For this reason, it is possible that there is a loss of color vision, the red diskette might not be

experienced as green or red, or there can be no conscious experience at all regarding the red diskette.

Gennaro does not give much detail on how such an unconscious experience would result, as he considers option 3 to be dependent on option 4: By ensuring a match between FOR and HOT, one guarantees a conscious experience. Levine constructs option 4 by attaching a demonstrative object to the content of HOT that refers to the first-order state. In this scenario, it is not possible for HOT to misrepresent and thus the match between two states is ensured. Levine, however, identifies two problems with this approach. Firstly, it remains possible for a HOT to be independently triggered, resulting again in a HOT that refers to an empty object as the first-order state. He asks: "Would that entail a sort of free-floating conscious state without a determinate character?" (Levine, 2001, p. 109). Secondly, it would fail to address the fundamental issue, namely the problem of overcoming the division of phenomenal labor. I agree with Gennaro that the second problem for option 4 just assumes the falsity of phenomenal labor rather than providing a substantial reason to deny option 4. Regarding Levine's first criticism to Option 4, Gennaro argues that we would not have a free-floating conscious HOT without a determinate character but just "an unconscious HOT without a target state" (Gennaro, 2004, p. 59). Accordingly, this would not result in a conscious experience, which is a plausible outcome that is in line with option 3, as Gennaro had argued. I think that Levine's criticism makes sense here and it can only be dealt with a theory that either takes FOR and HOT as jointly sufficient for consciousness to occur or with a self-representational theory of consciousness. In this respect, option 3 can be considered as the result of such a theory. Indeed, Gennaro himself defends some version of option 3 and 4 by connecting them explicitly to his wide-intrinsicality view, a

position considered by Rosenthal as a variant of self-representational theory. For this reason, I will deal with Gennaro's theory further in the next chapter when I discuss self-representational theories. For now, I will just add that even if Gennaro disagrees with Levine's criticisms to option 4, he explicitly agrees that these criticisms point to a problem in HOT theory: "it is difficult to make sense of entirely splitting off the lower-order state from the HOT" (Gennaro, 2004, p. 60). In my opinion, this acknowledgement closely mirrors the central critique made by Neander, Levine, and other critics of HOT theory, who focus on the conflict between overdetermination of HOT and division of phenomenal labor. I agree that HOT theory struggles to effectively resolve this issue, which constitutes the fundamental challenge that must be addressed. From this perspective, misrepresentation serves primarily as an illustrative tool, highlighting the specifics and implications of this deeper problem.

2.2.4 What-it-is-likeness

Intrinsic characterization draws support from the Nagelian definition of what-it-is-likeness. Supporters of intrinsic reading focus on “something” in “something-it-is-like” to take it as intrinsic quality of experience that can only be known subjectively from a first-person perspective (Weisberg, 2011, p. 411). Conversely, advocates of the extrinsic characterization of consciousness might give a different interpretation by focusing on the feature of for-meness and understand Nagelian phrase as something-it-is-like-for-me. They can emphasize that conscious experience is for an organism, and there is WIIL if there is something-it-is-like for a subject. This connection to the subject and his entire cognitive system can be cashed out functionally and in this way what-it-is-likeness can also be explained extrinsically.

Farrell considers the extrinsic characterization of Nagel's WIIL (which he defines as self-intimating reading of SOMETHING) to be non-standard. According to the standard reading, there being something it is like for S to be in M does not entail that S is aware of M. Janzen argues on the contrary. He starts with Nagel's formulation and attempts to create a formula of what-it-is-likeness from that. Nagel (1974) writes “An organism has conscious mental states if and only if there is something that it is like to be that organism – something it is like for the organism” (p. 436). Janzen (2011) argues that for Nagel there is a conscious mental state only if there is "something it is like for S to be in M at t" (p. 273). This requires that the subject S has a psychological integrity, which means S is in some way conscious of being in that mental state.

I partially agree with Farrell that Nagel's phrase does not explicitly apply Rosenthal's transitivity principle. However, it should not be seen as a notion in favor of an intrinsic quality without considering an awareness by the subject. To say that there is some kind of qualia and that this constitutes what-it-is-likeness without the presence of a form of self-consciousness is neither a neutral position nor aligned with Nagel’s definition. I think that Nagel's phrase should be seen as implicating some form of self-consciousness with what-it-is-likeness. Reaching a higher-order reading from this point is challenging because non-representational self-consciousness, already involved in mental quality, is another possible avenue. The concept of pre-reflective consciousness in phenomenological approaches, such as those discussed by Zahavi, serves as an example in this context.

2.2.5 Consciousness as the property of the subject

One way to deal with the problem of misrepresentation involves rejecting a specific interpretation of the relational structure within HOT theory. Explaining state consciousness through intransitive consciousness and higher-order thought does not mean that HOTs "transfer the property of being conscious from themselves to their targets; indeed, they don't induce any change whatever in those targets. Rather, they make their targets conscious because a state's being conscious consists simply in one's being suitably conscious of oneself as being in that state, and having a HOT is the right way of being conscious of oneself as being in a state" (Rosenthal, 2005, p. 185). Properly understood, HOT theory is not entirely relational, and consciousness is not a property of first-order state that is generated by and transferred from a higher-order state. As demonstrated in previous sections, many HOT theorists have defended a reading of consciousness as a property of (existing) mental states. However, this interpretation is mistaken, even as a version of HOT theory. Instead, transitivity principle reveals that consciousness is a property of the subject.

Berger presents two reasons supporting the view that consciousness should be understood as a property of the subject rather than of the state. Firstly, folk psychology does not exhibit any preference concerning this issue (Berger, 2014, p. 836). Secondly, treating consciousness as a property of state leads us to think of consciousness as if it is a property that HOTs somehow confer onto their intentional objects. It is not easy to understand how HOTs can do that and whether they represent a unique type of thought capable of doing so. This also introduces the complicated discussion regarding whether consciousness is a property constituted or transferred from the HOT or the subject to the first-order state. I think that folk-psychology might be even more sympathetic to view consciousness as a property of

the subject, thus avoiding the conceptual complications associated with attributing consciousness to mental states. Berger suggests a preferable approach to discussing consciousness, given that Rosenthal defines the transitivity principle as "one is in a conscious state only if one is aware of oneself as being in that state" (Berger, 2014, p. 837).

Rosenthal explicitly states that his account is not precisely relational: "A mental state's being conscious is not, strictly speaking, a relational property of that state. A state's being conscious consists in being a state one is conscious of oneself as being in. Still, it's convenient to speak loosely of the property of a state's being conscious as relational so as to stress that it is in any case not an intrinsic property of mental states" (Rosenthal, 2005, p. 211). It is debatable whether this loose characterization of relationality – intended primarily to contrast with intrinsicality – is genuinely helpful in addressing the issue of misrepresentation. Since HOT theory explicitly aims to explain state consciousness, it becomes crucial to clarify how consciousness as the property of the subject relates specifically to state consciousness. Should we understand this interpretation as an expression of transitive consciousness?

Clarifying the plausibility of such an interpretation significantly advances the discussion about misrepresentation. Under this interpretation, one becomes aware of oneself as being in a specific mental state, where the focus is on the subject rather than on the first-order state. This helps us to deal with the problem of non-existent first-order state since the subject always exists. When there is a HOT with a certain content, the subject is always the referent of this content, eliminating concerns regarding the potential non-existence of the subject. The mental states the subject is in become properties attributed to the subject. As we are conscious only under the

specific aspects through which HOTs represent us, we experience ourselves as being in mental states in relation to these higher-order representations. This also means that there is no real distinction between moderate and radical misrepresentation. In moderate misrepresentation, there is a subject in a mental state whose represented properties differ from those actually present in first-order sensation. In radical misrepresentation, there is again a subject in a mental state that differs from the first-order mental state. From the perspective of the subject, this difference is not substantial, as both types represent properties attributed to the subject. Thus, Berger (2014) states that "one is in a conscious state, whenever one is conscious of oneself in some respect or in a certain way means one is aware of oneself in respect of having the property of being in a mental state" (p. 838). In this sense, "being in a mental state" is the property of the subject and "red" is the property of the mental state.

Such a perspective initially appears relieving, particularly given the various problems discussed earlier. However, can all these issues genuinely be resolved with such a straightforward adjustment? Probably not. First, it remains unclear what kind of property redness represents. Is it the property of the first-order state or is it the property of the subject? If the latter, then the nature of HOT's content requires further clarification; specifically, one must explain how to speak of mental states having properties that are, in fact, properties of the subject. Second, a central motivation behind HOT theory is to explain state consciousness via transitive consciousness. The given characterization seems to support overdetermination of HOT, by emphasizing even more explicitly the role of the subject as primary while relegating everything else to a secondary status, seen as properties attributed to the

subject as the intentional object of HOT. In any case, the relation between the subject and HOT would play the fundamental role in such an understanding.

Gottlieb (2020) also builds on the idea that consciousness should be understood as a property instantiated by the subject rather than by mental states. He distinguishes between two models: the *instantiation model*, where "a conscious state is a mental state that instantiates the property of being conscious", and the *to-be-in model*, which describes a conscious state as "a mental state in virtue of which there is something it is like for (or to be) the subject of that state" (p. 6). Consequently, in cases of misrepresentation, neither the first-order state (M) nor the higher-order state (M*) instantiates consciousness. Gottlieb (2020) explicitly states, "The only thing that instantiates the property of being conscious are subjects" (p. 6). According to the instantiation model, only the subject (S) is conscious. The subject is phenomenally conscious by instantiating the right kind of HOT. In contrast, the to-be-in model holds that higher-order state M* is conscious, because in virtue of M* there is something it is like for S to be the subject of that state. Importantly, neither model attributes consciousness directly to the first-order state (M).

Recall that Farrell began with the premise 1 (M is conscious and S is not in M) and concluded with the contradiction in premise 8 (S is in M and S is not in M). Farrell argued that HOT theorist had to reject premise 4 (If there is an occurrence of WIIL associated with M, S is in M), but doing so would result changing the subject from the original reading of SOMETHING as used by Nagel. He further contends that the HOT Theorist equivocates between Nagelian standard reading of SOMETHING and a higher-order reading, making it impossible to reject the misrepresentation argument while simultaneously endorsing the awareness argument. In other words, the HOT theorist employs different readings across different

premises. Gottlieb responds to Farrell's criticism by rejecting premise 1. According to Gottlieb, one can affirm that the subject S is conscious (following the instantiation model) and that the higher-order state M* is conscious (following the to-be-in model). However, the first-order state M itself is never conscious; thus, according to HOT theory, the claim that “M is conscious” is not right.

Gottlieb formally articulates his position in terms of the transitivity principle. He distinguishes between two interpretations to which a HOT theorist might commit: *state-relational transitivity* and *self-relational transitivity*. In the former, a mental state is conscious only if its subject is aware of that state in an appropriate manner. In the latter, a state is conscious only if its subject is aware of oneself as being in that state in an appropriate way. While Farrell and numerous other critics assume that all HOT theorists commit themselves to state-relational transitivity, many HOT theorists actually ascribe to the self-relational transitivity, in which “the relata is the self”, who enters into an awareness-of relation (Gottlieb, 2020, p. 8). Gottlieb thus arrives at a refined formulation of the transitivity principle that operates at the level of subjects rather than mental states: "S is (non-introspectively) phenomenally conscious iff S is aware of herself as being in some (first order, representational) mental state” (Gottlieb, 2020, p. 9).

The emphasis on consciousness as a property of the subject also responds to Farrell's criticism that HOT theorists employ a self-intimating reading of SOMETHING, situating what-it-is-likeness (WIIL) with the presence of for-meness, and thereby requiring an awareness-of relation involving the subject. As Rosenthal (2011) argues, what-it-is-likeness is not a reliable common currency (p. 434). Nevertheless, interpreting WIIL in a way that necessitates the existence of its subject

is not fundamentally unconventional or off-topic.[13] While it would be incorrect to claim that Nagelian sense of WIIL explicitly implies an awareness-of relation or supports a higher-order explanation, it is reasonable to assert that WIIL coexists with a subject to whom it belongs. Moving beyond this neutral understanding toward either a HOT theory or a phenomenological account represents a further theoretical decision. However, it is not a standard way to understand Nagelian definition in terms of a qualitative property without the subject. The neutral reading recognizes qualitative experience, leaving the nature of its relation to the subject open to further theoretical specification.

2.2.6 Non-contingent relation between the lower and higher-order states

Denying the relational structure is not the only way to deal with the problem of misrepresentation. Another important option for higher-order theorists involves securing the representational relation between FOR and HOT to prevent the possibility of misrepresentation. Typically, this involves constructing a complex mental state (M) in which FOR and HOT are integrated as parts of a unified mental state where consciousness of M is explained through the HOT's representation of FOR. This approach combines elements from both first-order (FO) and higher-order (HO) theories, as consciousness is explained both through *being represented* and *representing*. The problem of non-existent first-order mental state is prevented because both FOR and HOT need to be present for consciousness of the whole mental state to occur. Their supporters also find their approach more advantageous because state consciousness becomes an intrinsic property of the integrated mental state itself.

[13] Remember that Farrell (2017) defended the idea that a reading of WIIL that necessitates the existence of its subject is nonstandard and it changes the subject.

I will use the term *higher-order intrinsic theories* (or self-representational theories, as a broader category) to refer to these modified higher-order representational theories. Such theories differ based on their understanding of the part-whole relationship among lower-order states, higher-order states, and the complex mental state as a whole, as well as on whether they emphasize self-representation through some form of reflexivity or the intrinsic nature of consciousness within the mental state. We will deal with different formulations of higher-order intrinsic theories in the next chapter, but a few preliminary comments are useful for concluding the current discussion. First, the concept of intrinsicality, which serves as a significant motivation for self-representational theories, is not thoroughly justified. In this direction, it remains unclear why intrinsicality should be accepted as a criterion that our theory of consciousness ought to fulfill. Second, we should decide how far we should prevent the possibility of misrepresentation. Once we consider it as an actual phenomenon, then occurrences of misrepresentation require adequate explanation. Self-representational theories might help to prevent radical misrepresentation, but it is worth recalling that within broader HO framework, the distinction between radical and moderate misrepresentation is typically denied. Furthermore, once we acknowledge the possibility of error as a natural outcome of any representational relation, it is not clear why more severe degrees of error should be considered impossible or distinct.

2.2.7 Causal role of first-order mental states

Before concluding this chapter and proceeding to an analysis of self-representational theories, it is crucial to clarify how Rosenthal conceives of the causal role of first-order mental states in generating higher-order thoughts. As previously discussed,

HOTs overdetermine the nature of conscious experience, leaving the precise explanatory role of first-order representations either unclear or altogether absent. At this point, it is essential to distinguish between *phenomenal role* and *causal role* of FORs. Phenomenally, the FORs do not determine the nature of conscious experience, which is completely dictated by the content of HOT. As we observed in the discussions on misrepresentation, first-order content can be misrepresented or completely overlooked by the content of HOT, which solely determines the subjective experience. However, this does not imply that first-order states do not have any causal role in the occurrence of higher-order thoughts.[14] Rosenthal (2005) explicitly states that "conscious mental states are mental states that [are companied by and often] cause the occurrence of higher-order thoughts that one is in those mental states" (p. 29). Accordingly, first-order states are often found among the causal factors triggering the HOTs, which in turn render these first-order states conscious.[15] Therefore, it would be misleading to conclude that FORs play no role whatsoever in HOT theory of consciousness. They possess an important causal role, and Rosenthal takes this causal contribution seriously (David Rosenthal, personal communication, June 2, 2025).

Nevertheless, Rosenthal rejects the idea that there is a causal requirement for a mental state's consciousness that this mental state has to cause the relevant higher-order thought (Rosenthal, 2005, p. 56). The requirement for a mental state's consciousness is simply that it needs to be accompanied by an appropriate HOT – not

[14] I would like to express my appreciation to Lucas Thorpe and David Rosenthal for pointing out this issue.

[15] The higher-order thought (HOT) makes its intentional object conscious. Frequently, the first-order state causing the occurrence of a HOT is also the object of that HOT – but this need not always be the case. There may be various first-order states contributing causally to the occurrence of a HOT, yet only one of them might ultimately be its intentional object. Alternatively, the intentional object of a HOT may be an entirely different first-order state that does not figure among the causal factors leading to the HOT, or it may be a non-existent state.

necessarily that there must be “some causal connection between that state and the accompanying higher-order thought” (Rosenthal, 2005, p. 27). It is one thing to claim that mental states, as the targets of accompanying HOTs, function as causal antecedents of those HOTs; it is quite another to assert that this causal relation between FORs and HOTs is a necessary condition for conscious experience. Indeed, various scenarios involving misrepresentation illustrate that individuals can form HOTs as being in certain mental states even when such mental states are absent. Cases such as dental fear demonstrate that a subject may form HOTs about “actual sensations of vibration and fear or about a nonexistent pain” (Rosenthal, 2005, p. 211). Consequently, the mental states that become conscious neither always cause nor necessarily need to cause the relevant HOTs.

Rosenthal explicitly rejects the idea that “for something to be the intentional object of an intentional state there must be some causal tie between them”, emphasizing instead that “that theory is no part of the higher-order-thought hypothesis itself” (Rosenthal, 1993, p. 162). Indeed, the causal connection between a higher-order thought and first-order state as its object may be quite remote or indirect. The object of HOT might only partially cause the relevant HOT, in conjunction with other mental factors. Furthermore, a higher-order thought might even be caused entirely by mental factors unrelated to the particular mental-state tokens it represents (Rosenthal, 1993, p. 162). Consequently, while investigating causal ties may occasionally illuminate the nature and formation of HOTs, it is largely irrelevant for HOT theory precisely which kind of causal connection exists between first-order states and their corresponding higher-order thoughts (Rosenthal, 2005, p. 211).

Rosenthal's views on the causal origin of higher-order thoughts show that merely attributing a causal role to first-order states does not fully address the problem concerning their precise contribution to conscious experience. Even if first-order states causally initiate higher-order thoughts – which is neither always nor necessarily the case – the phenomenal character of conscious experience is ultimately overdetermined by the content of HOT itself. Therefore, as our discussion of the misrepresentation problem illustrated, the exact explanatory role that first-order states play within conscious experience remains philosophically problematic.

2.3 Conclusion

In this chapter, we have examined how the problem of misrepresentation arises from a tension between the necessary condition, which states that we need both a first-order state and a higher-order state, and the sufficient condition, which holds that the mere presence of a HOT with appropriate representation is enough for consciousness to occur. Some critics find the underlying division of labor problematic, arguing that it relies on an intrinsic notion of consciousness which is presupposed rather than justified. Certain objections highlight the possibility of the required first-order state, posing a fundamental problem for the theory. Moreover, this issue raises concerns regarding the role of FOR in conscious experience and indicates that HOT may excessively or inappropriately determine the nature consciousness. Additionally, other critics argue that the explicandum of HOT theory differs significantly from other prevalent interpretations of what-it-is-likeness found in the literature.

While considering consciousness as an attribute of the subject addresses certain issues, it is not necessary to adopt this perspective. Ultimately, it underscored the significance of the transitivity principle; however, there are additional

fundamental aspects of a higher-order thought (HOT) theory that must be met. HOT theory is designed not only to address the idea that it is not meaningful to speak of consciousness of a mental state if we are entirely unaware of it, but also to clarify the mystery of phenomenal consciousness. The theory seeks to explain the distinction between conscious and unconscious mental states within a naturalistic framework with psychological terms. This is achieved by providing an explanation grounded in the relationship between unconscious states, which elucidates the concept of state consciousness. Thus, alongside the transitivity principle, several motivations remain equally important: (1) Distinguishing between conscious and unconscious mental states; (2) differentiating creature consciousness from state consciousness, with state consciousness explained as a form of intransitive consciousness through transitive consciousness; (3) providing a relational account of consciousness that demonstrates a viable alternative to intrinsic understanding of consciousness; (4) offering an explanation based on extrinsic, higher-order mental states, thereby distinguishing HOT theory clearly from first-order representational theories and self-representational theories. Viewing consciousness as an attribute of the subject can resolve certain issues, but it also risks oversimplifying the theory, potentially reducing its explanatory complexity to a one-dimensional perspective.

In this chapter, we focused on the nature of higher-order thought and explored how it overdetermines the conscious experience. In the next chapter, we will turn to self-representational theories, which emphasize a constitutive relation between first-order state and higher-order thought. We will address critical questions such as: What is the precise nature of the relation between FOR and HOT? Which aspect of this relational structure accounts for conscious experience – is it the specific relational structure between FOR and HOT, or is there something unique

about the nature of HOT itself? Alternatively, should we seek the essence of subjective experience within mental qualities themselves?

CHAPTER 3

SELF-REPRESENTATIONALISM

In Chapter 2, we examined several criticisms directed at HOT theory, most of which pertain to the relationship between the first-order representation (FOR) and the higher-order thought (HOT). These critiques insist that the relationship should be more constitutive or necessary. While I demonstrated that various forms of misrepresentation do not necessarily pose significant problems for HOT theory, it remains uncertain whether HOT theory is overall in a better position than self-representational theories to explain phenomenal consciousness. One might ask: Why not attribute consciousness to the entire mental state itself, making it both intrinsic to and a product of the relation between the first-order and higher-order states?

In this chapter, I will defend classical HOT theory against self-representationalism and argue that modifying it toward self-representationalism is not necessary. There is no compelling rationale for transforming HOT theory into a self-representational account unless one already adopts an alternative theoretical framework that inherently requires such a modification.

Theorists utilize various methodologies aimed at preserving some of the core principles of higher-order thought (HOT) theory while reassigning consciousness to the first-order state and integrating it into mental qualities. However, it is important to carefully assess the necessity and implications of this move within higher-order frameworks. In this chapter, I aim to demonstrate that HOT theory remains a strong candidate for explaining consciousness, without needing to adopt a self-representational approach.

There are various forms of self-representationalism, which will be analyzed throughout this chapter. *Reflexive theories* (or *pure self-referentialism*) explain self-representation by positing a reflexive mechanism that is directed from the mental state to itself as a whole. Accordingly, the mental state becomes conscious by virtue of its own self-directed structure. Such theories rely on a form of direct or pure self-reference to explain self-representation, rather than on a form of partial or indirect representation. Furthermore, they define self-representation as referring to the entire conscious mental state rather than as a representation of a self. Although reflexive theories come in various forms, most share a common difficulty: integrating them within a naturalistic framework and adequately explaining the phenomenon of self-causation. Specifically, one must clarify how this reflexivity arises, particularly if it involves a circular structure. Early forms of self-referentialism can be traced to Aristotle and Brentano, although whether these thinkers represent examples of pure self-referentialism remains debatable.

Higher-order intrinsic theories closely align with the classical HOT theory by explaining consciousness through a higher-order awareness (transitivity principle) but differ by considering this awareness an intrinsic feature of the conscious state. We can take Gennaro's *wide-intrinsicality* view as an example. He criticizes pure self-referentialism by arguing that such theories fail to adequately explain (a) what makes a state conscious and (b) how introspection works. To avoid problems arising from explaining circular self-representation, he suggests viewing first-order representation and the higher-order thought as integral parts of a global conscious mental state. These theories are categorized as higher-order intrinsic theories rather than self-representational theories because they stand close to the HOT family and

assert that some form of higher-order theory can effectively explain consciousness. They aim to refine HOT theory to more effectively address its critiques.

Self-representational theory (or *same-order monitoring theory*) explains consciousness as a form of self-awareness inherent to the conscious state. While endorsing the transitivity principle, it rejects the extrinsic nature of higher-order states, seeking instead to strengthen the relation between first-order representation and higher-order thought, by making this relation as constitutive and necessary. One way to establish this is considering the whole conscious state as a *complex* rather than merely a *sum* – a distinction Kriegel uses to differentiate self-representationalism from Gennaro's wide intrinsicality. For self-representationalism, it is necessary to have a first-order representation, a higher-order thought, and a proper relation between them for the whole conscious experience to occur. A psychologically real and structurally necessary relation between two mental states is required to prevent issues related to misrepresentation. This distinction, however, is open to debate. One might argue, following Gennaro, that conscious experience should indeed be considered a sum rather than a complex. Rosenthal, on the other hand, dismisses distinctions among different forms of self-representational theories entirely, classifying them under higher-order intrinsicality. He holds the view that all these versions rest on an unjustified assumption that consciousness must be intrinsic. Setting aside the ultimate validity of these distinctions, we can say that Kriegel's self-representationalism distinguishes itself from other theories by focusing on a psychologically and logically strong relation between the parts and the whole of conscious experience.

Embracing the transitivity principle ensures a relational understanding of consciousness, securing its place within a naturalistic framework. On the other hand,

consciousness can be viewed as an intrinsic property of the whole mental state, with both HOT and FOR becoming intrinsic components of this complex structure. According to proponents of this view, such integration prevents radical misrepresentation, ensuring qualitative features and consciousness remain inherently linked. While the phenomenal division of labor remains methodologically relevant to some extent, mental qualities and consciousness ultimately become interdependent. The notion of secondary representation, directed toward the mental state, is included in this explanatory framework, though it is not regarded as higher-order. Instead, both representations are considered integral elements within a single, unified mental state. Consequently, the explanation of consciousness becomes an explanation of self-representation, wherein these two distinct yet interconnected representations function collectively within one mental state.

Self-representationalism has other key features that warrant attention. We can introduce two mechanisms to explain how we can explain self-representation in a naturalistic framework. Firstly, self-representation can be understood as a form of *indirect representation*, where one part of the complex mental state indirectly represents the whole conscious mental state by directly representing another part of the complex. Secondly, one can draw an analogy between consciousness and attention structures, specifically using the peripheral/focal distinction, to explain the ubiquity of self. The idea is that the self is continuously present at the background of our conscious experience and is peripherally represented. Kriegel suggests that these mechanisms, unlike pure self-referentialism, can effectively account for self-representation within naturalistic boundaries (Kriegel, 2009).

The term “self-representational approaches” will be used as a general category encompassing all these theories. The term “self-representationalism” or

"self-representational theory" will specifically denote Kriegel's views, which have various versions. The reference of these terms should not constitute a problem, and it will be easily understood from the context of particular discussion. Certain differences are significant when addressing particular issues, and it is generally acceptable to regard all these theories collectively as variants of self-representationalism. Many of these theories aim to combine higher-order frameworks with one-level (representational or phenomenological) accounts.

3.1 Historical roots of self-representationalism

We have introduced three forms of self-representationalism. Let's consider the scenario in which I consciously see a brown ukulele. All self-representational approaches agree that whenever I consciously see a brown ukulele, I am also aware of my visual state of seeing the brown ukulele. Broadly speaking, self-representationalism rests on the idea that consciousness entails self-consciousness. (1) P*ure self-referentialism* posits that my visual state represents itself by being directed to itself as a whole. (2) *Higher-order intrinsicality* holds that my visual state contains a self-referential element, allowing the state to represent both the external object and itself. (3) *Self-representationalism* explains my conscious awareness through a unified internal state that represents both the external object (the brown ukulele with its perceived qualities) and itself (the subjective experience of me perceiving the brown ukulele). It can be further stated that there is a psychologically real and fundamental relation between one's mental state and its components (the representing part and the represented part).

I will begin by discussing pure self-referentialism, which aims to capture the central aspect of self-consciousness – how a mental state is directed to itself as a

whole. Accordingly, one is conscious of a mental state because "it has a content that turns back on itself" (Block, 2009, p. 1116). Aristotle and later Brentano (1874/1973) supported self-representationalism, which can be interpreted as early forms of pure self-referentialism. Although subtle differences exist, their arguments significantly influenced contemporary self-representationalists, particularly Uriah Kriegel. Next, the focus will be on higher-order intrinsicality, which criticizes pure self-referentialism on several grounds and remains closely aligned with HOT theory. Finally, I will explore self-representationalism. This approach shares some elements with HOT theory while incorporating insights from pure self-referentialism and phenomenological (non-representational) approaches.

3.1.1 Aristotle

The idea that a mental state represents itself while also representing the external world can be traced back to Aristotle (Kriegel, 2003; Zahavi, 2004). There is a consensus in contemporary philosophy of mind that Brentano was similarly motivated by Aristotle's thoughts on consciousness, particularly those presented in *De Anima*. Aristotle argues that every perception perceives both its external object and itself. Beyond simply offering an early form of pure self-referentialism, his criticisms against higher-order theory – specifically higher-order perception theory – remains relevant in contemporary discussions. Aristotle has two criticisms against higher-order perception theory: (1) It leads to an infinite regress, and (2) it requires a distinct modality for higher-order perception, resulting in an undesirable duplication of content.

3.1.1.1 Infinite regress

Attempting to explain consciousness of a first-order perception by invoking a secondary form of perception leads to a "vicious infinite regress" (Zahavi, 2020). If consciousness were explained by another distinct perception, this explanation would necessitate an endless series of perceptions. The only plausible way to halt such a regress would be to posit that at least one of these higher-order perceptions perceives itself. Therefore, it seems more reasonable to accept that first-order perception is inherently self-conscious.

When we perceive that we see and hear, it must be through sight or another sense that this perception occurs. However, it can be argued that the same sense which perceives sight will also perceive the color, which is the object of sight. This raises a dilemma: either there will be two senses perceiving the same object, or the sense of sight will perceive itself. Furthermore, if the sense perceiving sight were truly distinct, the series of senses would either continue infinitely, or one sense within that series would need to perceive itself. Therefore, it is more logical to conclude that this self-perception should apply to the first sense in the series[16] (Aristotle & Shields, 2016, p. 52).[17]

Aristotle, similar to modern self-representationalists, begins by criticizing higher-order theory and subsequently arrives at a form of self-representationalism. While contemporary critics of HOT theory primarily deal with the issue of misrepresentation, Aristotle focus on the problem of infinite regress. His opponent is *monitoring theory*, which can be seen as a reminiscent of contemporary higher-order perception theory of consciousness. Accordingly, a perception is conscious if it is also perceived by another state. Consider, for example, perceiving a red object.

[16] See Caston, 2002, p. 760 and Kriegel, 2003b, p. 111
[17] *De Anima III.2, 425b12*

Aristotle says, if there is a second sense perceiving the first-order sight of the red object, two possibilities emerge: either there are two senses about the red object, or the original sight of the red object also perceives itself. The former option is problematic, because then we would need to explain how the second sense becomes conscious. To explain this, one would need to postulate a third sense that would perceive the second sense, and that would go ad infinitum, resulting in an infinite regress. Ultimately, to halt the regress, one would need to posit an ancestor perception that is a perception of itself. Aristotle, therefore, argues in favor of the second option: The initial perception perceives both the object and itself. Rather than distinguishing one special ancestral perception from subsequent perceptions, Aristotle maintains that perception inherently involves self-perception from the outset. Thus, he begins with the premise that there is conscious perception – that whenever we perceive we also perceive that we perceive. The problem of infinite regress problem underscores the inadequacy of explaining conscious perception through distinct perceptions perceiving each other. Consequently, the only remaining viable explanation is that perception is inherently conscious.

I have simplified Caston's reformulation of Aristotle's "Regress Argument" into the following general form:

> 1. I perceive that I see and so have a perception of a perception.
> 2. Any perception of a perception is distinct from the earlier one.
> 3. Whenever one perceives, there will be an infinite chain of perceptions, each new one being a perception of the earlier one.
> 4. But it is impossible to have an infinite chain of perceptions.
> 5. In a finite chain of perceptions, the last member of the chain will be a perception of itself.
> 6. There is no more reason for some perceptions (last perception) rather than others (first perception) to be perceptions of themselves.
> 7. The first perception is already a perception of itself. (Caston, 2002, p. 773)

Caston argues that Aristotle's argument relies on two critical assumptions (especially the derivation of premise 3 from premises 1 and 2), which is otherwise invalid: (i) A perception of a perception is also of what the earlier perception is of; (ii) whenever we have a perception, we have a perception of that perception. The first assumption introduces a transitive notion of perception, which leads to an infinite chain of perceptions (Premise 3), but only if we exclude the possibility of loops. Possibility of an infinite chain arises specifically from this transitive mechanism established by the first assumption: If perception *a* represents perception *b*, and perception *b* represents perception *c*, then perception *a* represents perception *c* by virtue of representing perception *b* (Premise 1). The regress continues indefinitely only if loops are prohibited (Assumption 1).

Second assumption requires us to posit another perception at each successive level (Caston, 2002, p. 774). The idea of perceiving a perception leads to infinite regress only if one presupposes that the second-order sight must itself be conscious. Otherwise, it is possible to argue that the higher-order sense is unconscious – a move used by higher-order theories, specifically by higher-order perception theory, wherein an unconscious second-order perception renders the first-order mental state conscious by perceiving it. However, Aristotle excludes this possibility. Although the precise nature of his position remains subject to debate, a plausible interpretation is that Aristotle's assertion "whenever one perceives, one perceives that he perceives" should be understood conditionally, rather than attributing him the stronger claim that all mental states are conscious. Caston notes that this broader assertion can indeed be found in Aristotle's works (Caston, 2002, p. 775). Brentano reflects a similar position, explicitly rejecting the existence of unconscious mental states. He argues that acknowledging such states would imply that consciousness emerges from

the unconscious, which he finds untenable (Zahavi, 2020, Ch. 2). Kriegel finds Brentano's position that "all mental states are conscious" controversial, though he suggests it could be defended as the reformulated claim "all experiential mental states are conscious" (Kriegel, 2003, p. 122). Zahavi contends that explaining consciousness through unconscious higher-order states is problematic. Moreover, he asserts that merely stating that perception is conscious of itself is insufficient and such a claim raises additional complexities if one adheres to a reflection theory (Zahavi, 2020, Ch. 2).

Most self-representational theories attempt to preserve a form of the transitive concept of consciousness by explicating it through a representational relationship. At the same time, they aim to characterize consciousness as an intransitive attribute of the mental state, ultimately treating it as an intrinsic and/or reflexive property of the entire mental state. Kriegel, for instance, explains the intransitivity of consciousness through an analogy to *peripheral attention*. Aristotle, on the other hand, employs transitivity as a methodological assumption to illustrate the possibility of infinite regress. However, the form of intransitive consciousness Aristotle arrives at in the conclusion – the self-perception inherent in the initial perception – is weaker compared to contemporary self-representationalism. According to Caston (2002), Aristotle holds the view that the perception "perceives that it sees in virtue of the nature of perception more generally" (p. 779).

Caston's interpretation of Aristotle presents a modern form of self-representationalism, proposing an intrinsic understanding of consciousness that is also relational at the same time. Unlike Rosenthal's approach, the distinctions between intrinsic versus extrinsic and relational versus nonrelational are not opposing alternatives in Aristotle's framework. Aristotle's position "preserves our

intuitions about the intrinsic nature consciousness, while also providing informative explanation in terms of intentionality" (Caston, 2002, p.781). As we have seen, an important motivation behind Aristotle's position is to address and avoid the problem of infinite regress. Aristotle prevents infinite regress by denying there are distinct mental states when we have conscious perception. Rosenthal, by contrast, avoids infinite regress by denying assumption (ii) – the idea that whenever we have a perception, we have a perception of that perception. Examining Aristotle's arguments and assumptions thus sheds light on debates between HOT theory and modern self-representational approaches.

3.1.1.2 Duplication of content

Another problem Aristotle identifies with higher-order perception theory is that it leads to the double perception of the same object. Caston formulates *the duplication argument* as follows:

> 1'. Given that we perceive that we see, there will be a single perception of our seeing and the azure of the sky.
> 2'. We perceive that we see either (i) by means of a distinct, new perception or (ii) by means of the initial perception, the seeing.
> 3'. Either (a) there will be two perceptions of the same thing (namely, azure), or (b) the one perception will also be of itself.
> 4'. But there are not two perceptions of the same thing. (Caston, 2002, p. 771)

I will apply this reasoning to my earlier example of my perception of a brown ukulele: According to the first premise, there is a single perception encompassing both its brownness and my act of seeing it. The second premise specifies that this single perception must occur either via a distinct perception or due to the intrinsic nature of initial act of seeing. Aristotle rejects the first option (distinct perception) on phenomenological grounds, as this would imply that we experience something akin to "double vision" (Caston, 2002, p. 772). Therefore, to avoid duplicating the object

of perception (sight and brownness), we should conclude that perception inherently perceives itself.

The duplication argument provides additional support for assumption (ii), as employed in the infinite regress argument. This argument closely parallels Rosenthal's critique of higher-order perception theory. Rosenthal emphasizes that there is no special modality for a higher-order perception, which prevents us from accepting that the higher order mental state is perceptual in nature. Accepting such a modality would require acknowledging additional perceptual content specific to this higher-order perception. Aristotle emphasizes concerns related to the duplication of perceptual content by focusing on the quantity of perceptual contents. In contrast, Rosenthal's critique focuses primarily on qualitative concerns: If consciousness were attained through a distinct perception, there should be an additional qualitative content associated specifically with this secondary perception, which phenomenologically we do not find.

Aristotle does not entertain the possibility, as Rosenthal does, that consciousness results from a distinct thought. Instead, he argues that the perception should also be of itself and explicitly denies the formulation of HOP theory. This does not imply that Aristotle endorses some form of perceptual self-representationalism. For Aristotle, there is no postulated mechanism that is perceptual in nature; rather, "the perceptual system sees in virtue of its visual part" (Caston, 2002, p. 779). As discussed in the previous section, consciousness understood as reflexivity is generally present in normal perception, but this does not mean that the perceptual mental state somehow perceives itself. This reflexive feature can similarly apply to thinking, but it is not an intrinsic feature of thought.[18] It

[18] We can arguably say that in HOT theory, while higher-order thoughts have such a reflexive feature through inner awareness, this does not apply to all thoughts.

is beyond the aim of this thesis to decide on correct interpretation of Aristotle. Nevertheless, suggesting a potential Aristotelian stance on conscious thought may be dialectically beneficial, considering the ongoing discussion of his views on self-representationalism: Aristotle might hold that thinking is conscious by virtue of the thinking in general, without necessitating any higher-order thought being responsible for consciousness of first-order thoughts, as this would lead to an infinite regress. Aristotle's argumentation works against both HOT and HOP theories, as the main problem of infinite regress persists independently of whether the second-order mental state is considered perceptual or thought-like. Arguably, his position toward consciousness of perception and thought would be similarly structured.

It is debatable whether Aristotle's view can be accurately characterized as a form of pure self-referentiality. However, it clearly represents a version of self-representationalism based on self-reflexivity. Accepting Caston's interpretation opens new ways to engage with contemporary discussions surrounding HOT theory, particularly by demonstrating how one can adopt a relational account while maintaining an intrinsic notion of consciousness through a self-representational approach. Aristotle does not aim to give an intrinsic account of consciousness merely to justify a preconceived assumption that consciousness must be intrinsic. Rather, his conclusion that consciousness should be intrinsic emerges specifically as a response to the problems of infinite regress and duplication of content. Rosenthal criticizes self-representational theories for preserving what he sees as an unjustified assumption that consciousness should be intrinsic. However, Aristotle's rationale for considering consciousness as intrinsic differs from this characterization.

3.1.2 Brentano

Brentano also presents a form of self-representationalism by arguing that our perception of an object is united with its own awareness under one unified mental state:

> In the same mental phenomenon in which the sound is present to our minds we simultaneously apprehend the mental phenomenon itself. What is more, we apprehend it in accordance with its dual nature insofar as it has the sound as content within it, and insofar as it has itself as content at the same time. We can say that the sound is the primary object of the act of hearing, and that the act of hearing itself is the secondary object.[19]
> (Brentano, 1874/1995, p. 98)

Accordingly, we do not have two separate mental states (perception of x and consciousness of the perception of x). Rather, our perception and consciousness of that perception together form a singular unified mental state, in contrast to higher-order thought theory. Within this unified mental state, we have awareness of two objects. These two aspects do not constitute separate mental states but rather they are two objects within a single act of perception.

Brentano utilizes the arguments from infinite regress and duplication of content in support of his position. Smith's interpretation is the typical way to relate Brentano to Aristotle: "Reviving an ancient argument, Brentano reasoned that the secondary presentation cannot be a second presentation, a reflection or judgement upon the primary presentation. For that would lead to an infinite regress […]" (Smith, 1986, p. 149). For Textor (2017), Brentano's main position hinges on the idea that all mental acts are conscious, and the possibility of infinite regress poses a challenge that must be addressed to uphold this view (p. 55). Otherwise, one would be inclined to defend a form of HOT theory, which would represent the remaining alternative to prevent infinite regress. Such a solution would obviously be

[19] See Gallagher & Zahavi, 2020, p. 60 for a discussion.

unacceptable by Brentano, because HOT theory allows for unconscious higher-order states. Although infinite regress remains a significant issue Brentano must address, Textor emphasizes duplication of content plays a more substantial role in shaping Brentano's positive view.

We have already discussed the duplication argument in the previous section. If there were two different mental acts – one representing a mental quality and another representing consciousness of this representation – this would result in representing the mental quality twice. For example, when consciously perceiving a red apple, the quality of redness would be represented again, alongside the consciousness of the mental state, because the act of conscious-making differs from the first-level perception. However, the inner experience suggests that the mental quality is represented only once. In support of this view, Textor (2017) draws upon Anthon Marty's (2011) argument[20] that the representation of redness and the representation of this representation are intimately connected within a single mental phenomenon, with the result that the mental quality is represented once (p. 55). Textor outlines the general structure of the argument as follows:

> (P1) Assume for reductio that S's perceiving S perceiving x is distinct from S's perceiving x.
> (P2) S's perceiving S perceiving x is directly of S's perceiving x as well as indirectly of x.
> (C1) Hence, when S perceives S's perceiving x, x is presented twice: once in S perceiving x, and a second time in S's perceiving x.
> (P3) There is no double presentation of x when S perceives S's perceiving x.
> (C2) Hence, S's perceiving S perceiving x is the same act S's perceiving x.
> (Textor, 2007, p. 56)

In this formulation, Brentano and Aristotle regard P3 as a phenomenologically justified fact. Therefore, it is important to demonstrate that P2 is more plausible than P1. According to Brentano, "it seems plausible that the tone must be contained not

[20] Marty, A. (2011). Deskriptive Psychologie. Würzburg: Königshausen & Neumann.

only in the hearing, but also in the simultaneous presentation of the hearing as something presented" (Brentano 1874/1973, p. 171, my translation).[21] However, this alone is not enough to justify choosing P2 over P1, because an opponent can argue that "awareness provides us only with access to mental acts and not also with access to their objects" (Textor, 2017, p. 56). Therefore, it is crucial to clarify how two mental acts can constitute a single unified state, an analysis Brentano explores further in his writings. Accordingly, there is a single mental phenomenon, which can be conceptually differentiated into two different acts once we examine them in relation to different objects: One act pertains to a physical object, while the other one is related to a mental act. Hence, the differentiation is a matter of conceptualization, analogous to how a spatial point "a" – though indivisible as it does not have any parts – can be relationally conceptualized in two different ways: "the left neighborhood of point b" and "the right neighborhood of point c" (Textor, 2017, p. 56).

Brentano's views have significantly influenced the development of self-representational theories, providing a viable alternative to higher-order representational theories. Kriegel characterizes a version of his theory explicitly as neo-Brentanian. However, such accounts face criticism from both by HOT theorists and phenomenologists. Rosenthal argues that self-representational theories are motivated by the unjustified claim that consciousness should be intrinsic. Zahavi finds over-usage of Brentano in contemporary self-representationalism unnecessary, because it is not the "only one-level game in town" (Zahavi, 2004). Gennaro considers Brentano's framework as a form of pure self-referentialism, wherein self-

[21] The original text by Brentano (1874) is as follows: "Auch das scheint einleuchtend, daß der Ton nicht bloß im Hören, sondern auch in der gleichzeitigen Vorstellung des Hörens als vorgestellt enthalten sein muß" (p. 171). Textor translates "Vorstellung" as "presentation" rather than "representation" and I sticked to this decision. (See Textor, 2007, p. 56)

representation is understood as the mental state turning back on itself and situates his wide-intrinsicality view as a middle ground between pure self-referentialism and (extrinsic) HOT theory. However, numerous interpretations of Brentano exist, many of which explain the relation between two mental states to demonstrate how they unify into a single mental state. If a coherent Brentanian explanation for self-reference can be given, this position could be seen as an early form of self-representationalism, closely aligned with Kriegel's perspective. We have previously argued that the differentiation between acts can be understood as merely conceptual. Pursuing this line of thought allows us to further clarify self-representation by examining the relationship between a unified mental state and its parts. There are additional indications in Brentano's writing suggesting such a route:

> Every mental act is conscious; it includes within it a consciousness of itself. Therefore, every mental act, no matter how simple, has a double object, a primary and a secondary object. The simples act, for example the act of hearing, has as its primary object the sound, and for its secondary object, itself, the mental phenomenon in which the sound is heard." (Brentano, 1874/1973, p. 119)[22]

Kriegel (2009) attributes this idea to Aristotle's *Metaphysics*: "knowing, perceiving, believing, and thinking are always of something else, but of themselves on the side" (p. 14). Considering self-representation as the secondary object (Brentano) or maintaining that any perception inherently includes consciousness of itself on the side opens new pathways for explicating the self-representational idea of consciousness. Rejecting higher-order monitoring consciousness – that consciousness emerges from the relation between two mental states – does not necessarily entail adopting a circular view of self-representation. It does not necessarily lead to mysticism either, where perceptual states are assumed to be

[22] See Kriegel, 2003b, p. 119 for another discussion of this passage.

inherently conscious of themselves without further explanation. Perception could instead be indirectly aware of itself, indicating that the primary object (such as a red object) and the secondary object (the perceptual act itself) need not be placed on equal footing. Consciousness may still be representational, but the representational mechanisms might differ from those involved in ordinary direct perception. Thus, consciousness of the mental state could be considered secondary, peripheral or present in the background of any perceptual act, while the object of perception remains primary or focal. Modern forms of self-representationalism are distinguished from pure self-referentiality precisely through their ability to explain how self-representation is possible, and the varying mechanisms underlying self-representation lead to distinct formulations of self-representational theories (Kriegel, 2009, p. 14).

Highlighting Brentano's inherent explanatory mechanisms does not necessarily confirm that Brentano was a self-representationalist in the strict sense. His assertion that all mental states are conscious complicates categorizing him neatly within either a physicalist framework or Husserl's phenomenological framework. Zahavi notes that Brentano correctly criticizes reflection theories, asserting that "our intentional act does not need to await a secondary act of reflection in order to become self-aware" (Zahavi, 2020, Ch.2). Reflection theories take self-awareness as a reflective relation between two mental acts, resulting in an infinite regress. However, Brentano's view that the perception of an object also includes consciousness of the entire mental act leads to a similar problem, because self-consciousness again becomes the outcome of a secondary reflection, albeit an unusual one. This secondary reflection does not pertain solely to the perception of an external object (e.g., the color red), but rather encompasses the entirety of the mental state, including

the consciousness of the co-consciousness already present in the initial perception. Consequently, this approach: (i) continues to make use of reflection, and (ii) does not properly explain intrinsic self-consciousness already present in the perceptual act. Zahavi thus argues that Brentano's position does not represent a genuine form of self-representationalism, but rather an unusual variant of reflection theory. Accordingly, Brentano should have adopted a more radical position by completely rejecting any use of reflection. Without doing so, Brentano's argument loses explanatory power. If self-consciousness is already present at the first-order level, how could a secondary act of reflection, even if it is a form of self-reflection, sufficiently explain this original self-awareness? For Gennaro, this is a form of pure self-referentialism and it is problematic at a more fundamental level due to the difficulties inherent in explaining pure self-reflection within a naturalist framework.

3.2 Contemporary pure self-referentialism and its criticism

Pure self-referentialism (PSR) is based on the identity of the representing and represented states. Gennaro employs PSR methodologically as a radical alternative to Rosenthal's HOT theory, subsequently presenting his wide-intrinsicality view (WIV) as a moderate stance between classical HOT theory and PSR. Similarly, Kriegel utilizes forms of PSR to differentiate among various self-representational theories. I will follow these established categorizations and regard PSR as a viable form of self-representationalism.

Kriegel (2006) follows a distinction between higher-order monitoring theories (HOMT) and same-order monitoring theories (SOMT) in the following way:

> (HOMT) A mental state M of a subject S is conscious iff S has a mental state M*, such that (i) M* is an appropriate representation of M, and (ii) there is no constitutive relation between M and M*.

> (SOMT) A mental state M of a subject S is conscious iff S has a mental state M* such that (i) M* is an appropriate representation of M, and (ii) there is a constitutive relation between M and M*. (p. 145)

Same-Order Theories explain consciousness through a representational account based on the notion of *being represented*. They concur with first-order representationalists and higher-order representationalists that consciousness must be understood in terms of representational activity. Same-order theorists differentiate their view by positing a constitutive relation between the represented and representing states. According to same-order theories, this relationship is necessary for consciousness, whereas higher-order thought theories regard it as contingent, suggesting that an unconscious mental state can exist without being represented. Exploring the precise nature of this fundamental and constitutive relation has led to various formulations of higher-order and same-order theories. One notable possibility is that the relation is one of identity. As previously discussed, this idea traces its historical roots back to Aristotle and Brentano. Kriegel (2006) formulates this option as follows: "(SOMT1) A mental state M of a subject S is conscious iff S has a mental state M*, such that (i) M* is an appropriate representation of M, and (ii) M=M*" (p. 146).

We have seen how this picture might present the most plausible alternative if one wishes do avoid both the view that all mental states are always conscious and the position that conscious states become conscious only through being represented by another distinct mental state. The first position does not retain the idea that conscious states are represented mental states, while the second position faces all the issues associated with HOT theory. A pure self-referential account can effectively explain how a mental state is conscious by both representing and being represented. This is

only possible if the mental state represents itself – that is, if the relation between the representing and represented state is one of identity.

One significant challenge with the relation of identity, whether in Brentanian form or contemporary self-referentialism, is explaining its compatibility with a naturalistic and physicalist framework. A notable advantage of HOT theory – beyond its intrinsic theoretical strengths – is its suitability for such a naturalistic explanation. HOT theory accounts for consciousness through a representational relation, which is a familiar philosophical tool that is broadly applicable in various philosophical contexts. In the HOT paradigm, there is a coherent explanatory transition from physical states to mental states (via intentionality) and subsequently to conscious states (through appropriate representational relation between two mental states). A self-representational theory similarly requires a conceptual tool capable of ensuring naturalistic compatibility.

The utilization of naturalism-friendly tools in this context is generally experimental. There are biological as well as mathematical or computational approaches aimed at explaining how self-reference possible. Modern phenomenological approaches integrate insights from psychology, while embodied approaches draw from biology and evolutionary psychology. However, these latter two frameworks do not explicitly argue for self-reference or self-representation; instead, phenomenological approaches emphasize pre-reflective subjectivity, and embodied approaches stress a non-reflexive, intimate relationship with the environment. Analyses employing such tools typically depend on their respective frameworks, though fruitful interactions can emerge from attempts to align theories of consciousness with physicalist principles. For instance, Kriegel (2009) employs the notion of peripheral attention, while Kriegel (2009), Coleman (2015) and Brown

(2025) seek to explicate self-representation through forms of indirect representation. Criticisms directed at these applications are not exclusively problematic for pure self-referentialism; rather, they equally affect self-representationalism more broadly and thus do not constitute a specific weakness unique to PSR.

Returning to the primary issue, it remains "unclear how the capacity for self-representation is supposed to be implemented in a natural, physical system" (Kriegel, 2003, p. 481). Accordingly, the idea of self-representation is seen mysterious, prompting skepticism about the feasibility of any physicalist explanation. A common physicalist assumption is that no physicalist account of mental representation could adequately clarify self-representation. This skepticism is influenced by Brentano, who maintains that conscious states are self-representational but not physical. Modern self-representational theories, described as neo-Brentanian, attempt to integrate two principles: (a) conscious states are self-representational, and (b) conscious states are physical. To reconcile these principles, proponents must explain how self-representational states can be instantiated physically, thus employing conceptual tools typically applied within physicalist frameworks (Kriegel, 2003, p. 481). In other words, it must be demonstrated how a mental state can render itself conscious by being directed upon itself.

Gennaro (2006) follows Kriegel's working definition and argues that numerous issues render PST implausible (p. 222). First, he claims that PSR does not really attempt to explain consciousness but merely illuminates the structure of a conscious mental state. Gennaro attributes this issue to non-physicalist assumptions shared among PSR theorists: There is a tendency to reject reductive explanations of consciousness[23] and deny the possibility of unconscious mental states altogether

[23] See Smith, D.W., 2004, "Return to Consciousness," in *Mind World: Essays in Phenomenology*

(Brentano, 1874/1973). It is not properly explained how M* is identical to M in a relation where M* is directed to M. Furthermore, it is problematic to explain the status of M* and M in such a relation in terms of being conscious/unconscious: How can an unconscious M* render M conscious, when M* is identical to M? One either must assert that M is already conscious or develop a modified version of HOT theory to explain such self-representation. Proposals that regard M* and M as parts of some meta-state, or arguments suggesting that "M becomes conscious when it acquires a particular self-referential content" (Kriegel, 2012, p. 106) exemplify such modified approaches and thus cannot be categorized as forms of pure self-referentialism.

Secondly, PSR is problematic insofar as it makes use of an attention mechanism that is both world-directed and inner-directed. Gennaro argues that such dual-directional attention is not plausible since attention is typically directed either toward objects in the external world or toward inner mental states – but not both simultaneously. Ordinarily, we would consider the state M to be world-directed and the state M* to be inner directed. The question thus arises: How could M* also be world-directed at the same time? Gennaro argues this is not phenomenologically justified, as we typically shift our attention between outer-directed consciousness and introspective consciousness. PSR fails to plausibly explain the difference between these two types of attention. If one starts with an explanation of M by arguing that M itself is both outer-directed and inner-directed, this will imply that M already is conscious in some way, thereby undermining the goal of explaining self-referentiality through attention. Such a move would necessarily involve distinctions between parts and wholes, effectively resulting in a modified HOT theory.

and Ontology, Cambridge: Cambridge University Press; and Thomasson, A., 2000, "After Brentano: A One-Level Theory of Consciousness," *European Journal of Philosophy* 8: 190–209.

Third criticism by Gennaro is that PSR does not offer a plausible account of introspection. One important motivation to support HOT theory is its straightforward capability to give different explanations for unconscious mental states, conscious mental states and introspective conscious states: A mental state without an accompanying HOT is unconscious; the presence of an appropriate HOT leads to conscious experience, and the addition of a third-order HOT accounts for introspective consciousness. By contrast, the reliance on the identity between M* and M in PSR renders such differentiation problematic.

The final point is phenomenological. An important motivation for PSR and self-representationalism in general is the phenomenological intuition that mental states are intrinsically conscious. Thus, proponents argue that it is preferable to account for consciousness with a structure in which both the representing and represented mental states are conscious. In other words, within PSR, the state M* (and, analogously, the meta-state in self-representationalism) should itself be conscious. However, relying on phenomenological considerations to shape philosophical arguments about consciousness can be problematic, as it may lead philosophers to selectively utilize conceptual tools merely to fit specific phenomenological descriptions. Moreover, Gennaro contends that our phenomenology of consciousness does not necessarily support PSR or self-representationalism. Specifically, it is not inherently plausible to say that the representing state M* is conscious rather than unconscious. The idea that M* should be conscious arises from the assumption that inner awareness, as the feature responsible for self-representation, should itself be conscious. Gennaro (2012) argues that accepting the existence of such inner awareness does not entail that this awareness is phenomenologically revealed (p. 114). Indeed, HOT theory

accommodates the idea of inner awareness by suggesting that HOTs, which are inherently unconscious, are responsible for inner awareness. Thus, we require additional moves to conclude that M* is conscious, and these depend on the particular assumptions used in the specific theoretical framework. In first order representationalism, first-order states are already taken to be conscious. In self-representationalism, it is argued that first-order states become conscious through representation, yet the representing mechanism itself must also be conscious.

Gennaro contends that many PSR accounts can be understood as variations of wide-intrinsicality, which is the modified HOT theory he advocates. Numerous theories ultimately resort to discussing the relation between parts and wholes to account for the reflexive relation between mental states, thus raising doubts about whether such theories can genuinely be categorized as pure self-referentialism. Gennaro (2012) references Natsoulas, Smith, and Kriegel to support this claim (p. 115). He argues that even Brentano could be understood as a defender of some form of wide-intrinsicality: "The presentation which accompanies a mental act and refers to it is part of the object on which it is directed" (Brentano, 1874/1973, p. 99). Accordingly, the secondary representation is integral to the entire conscious state: M* is not identical to M, but it is part of the whole conscious state. Natsoulas (1993) states:

> Also, each conscious mental-occurrence instance includes as an intrinsic dimension of it a noninferential conceptualizing awareness of the mental occurrence instance… Not only is a conscious mental-occurrence instance presented and directly apprehended in its qualitativeness, but also there is, as part of its occurrence, awareness of it as this (a mental-occurrence instance) or that (something else). (p. 117)

Smith (1989) articulate a similar viewpoint:

> A conscious mental event must include an *occurrent* awareness of its own transpiring; the subject must *experience* the event as it transpires. Furthermore, this awareness must be an occurrent part of the given mental

> event itself, for reasons that will emerge. Thus, it is an *inner* awareness, a constituent awareness of a part of the subject's own stream of consciousness. (p. 81)

Pure self-referentialism (PSR), although a subset of self-representationalism, does not constitute its strongest form due to inherent issues related to identity. Gennaro rightly notes that Kriegel utilizes a form of indirect self-representation instead of direct self-representation, thereby distancing himself further from PSR and placing his theory closer to Gennaro's modified HOT theory. It can be further asserted, as Rosenthal contends, that all forms of self-representationalism may be considered adaptations of HOT theories. Despite ongoing debates regarding distinctions among various self-representational theories, there is consensus on the idea that pure self-referentialism does not represent the most robust form of self-representationalism. For this reason, it is better to consider Gennaro's wide intrinsicality and Kriegel's self-representationalism as the most powerful alternatives to Rosenthal's HOT theory.

3.3 Higher-Order Intrinsic Theories

The discussions regarding the problem of misrepresentation in Chapter 2 indicated that securing the connection between first-order and higher-order mental states may be necessary for those wishing to maintain higher-order theory rather than adopt first-representationalism. One prominent option is to argue that the conscious state represents itself while representing the object in the external world. However, as previously demonstrated, this approach struggles to explain pure self-representation within a physicalist and naturalist framework. Additionally, it leads to further local problems related to attention, introspection and the phenomenology of conscious experience. Certainly, one can dive into the phenomenological literature and explain

self-consciousness through pre-reflective self-awareness, but this would be a radical departure from the representational framework commonly employed in physicalist and functionalist theories of consciousness. There are lots of experimental alternatives attempting to integrate the tools from various philosophical disciplines into the philosophy of consciousness to explain pure reflexivity. However, analysis of such tools is beyond the limit of this thesis. Therefore, a viable option is to maintain a position close to classical HOT theory while incorporating insights from pure self-referentiality and phenomenology. This is precisely what theorists such as Gennaro (wide-intrinsicality view), Kriegel (self-representational theory) and Van Gulick (higher-order global states theory) attempt by asserting that some form of self-awareness is intrinsic to the conscious state. At the same time, they seek to elucidate this intrinsic self-awareness without abandoning key principles of extrinsic HOT theory, particularly the transitivity principle: A mental state is not conscious if one is in no way conscious (aware) of that state (Rosenthal, 2005, p. 4). They employ the *being represented by a HOT* relation to explain how a mental state becomes conscious.

In the subsequent analysis, I will particularly examine the views by Gennaro and Kriegel. Gennaro's approach represents higher-order intrinsic view, while Kriegel advocates self-representational approach. These views differ essentially in their proximity to traditional HOT theory: Gennaro's view remains closer to traditional HOT theory, whereas Kriegel integrates elements from pure self-referentialism, phenomenology and first-order representational theories of consciousness.

Gennaro's *wide intrinsicality view* (WIV) makes use of the notion of *meta-psychological thought* (MET) rather than higher-order thought (HOT). The first-

order state (M) and the meta-state (MET) constitute the *complex conscious state* (CMS). This CMS becomes conscious when an unconscious MET represents M.

> WIV: A mental state M of a subject S is conscious if and only if S has a suitable (unconscious) metapsychological thought M* (=MET), directed at M, such that both M and M* (MET) are proper parts of a complex, conscious mental state CMS. (Gennaro, 2006, p. 222)

This structure allows for a form of consciousness that is intrinsic to the complex state while still relying on a relational explanation involving its constituent parts. WIV maintains compatibility with naturalistic frameworks by positing an unconscious MET, thus avoiding a problematic notion of mysterious self-referential consciousness.

The main reason why Gennaro advocates for his WIV over HOT theory is his claim that "consciousness seems to be an intrinsic property of conscious states", and his theory better accumulates this aspect (Gennaro, 2006, p. 222; Gennaro 2012, p. 57). He notes that even Rosenthal acknowledges that we should preserve this intuition if possible (Rosenthal, 1986, pp. 331, 345). For Gennaro, consciousness should not be understood as a relational property, akin to spatial relations such as *being to the left of*, but rather as a part of the mental state itself (Gennaro, 2006, p. 222). Even if we accept Rosenthal's claim that consciousness need not reveal all phenomenological aspects about itself, Gennaro argues that Rosenthal's position does not sufficiently demonstrate why consciousness cannot be intrinsic. While Rosenthal contends that an intrinsic understanding of consciousness is unanalyzable and uninformative, Gennaro rightly criticizes him by arguing that Rosenhal's position is not strong and leaves us in a "false dilemma" (Gennaro, 2012, p. 55). According to this dilemma, one must either embrace a non-intrinsic account of consciousness or settle for an uninformative explanation based on mistaken phenomenological intuitions about consciousness: " … if consciousness is what

makes a state a mental state, consciousness will not only be an intrinsic nonrelational property of all mental states; it will be unanalyzable as well" (Rosenthal, 1986, p. 341). Contrary to Rosenthal's position, Gennaro argues that there are compelling reasons to accept the intrinsic nature of consciousness and adopting this stance can yield an explanatory account.

Modified HOT theories construct the conscious mental state as a unified complex state, subsequently disintegrating this complex into constituent parts to explain how one part renders another part (or the whole mental state) conscious through representational relations. Determining the precise nature of these parts (whether they are proper or improper), the nature of the complex (whether consciousness is simply the sum of its parts or a holistic entity above its parts), and the exact relationship between the parts and the complex (such as identity relations and indirect self-reference) lead to varieties of self-representationalism. Rosenthal refrains from engaging extensively in these local discussions about parts and wholes. For him, there is no non-arbitrary way to distinguish the modified HOT theories from his extrinsic HOT Theory (Rosenthal, 1986, p. 345). Moreover, self-representational theories do not really solve the problem of misrepresentation: Even if all parts of the conscious mental state were conscious, these theories would remain vulnerable to the same regress and circularity issues that critics highlight against classical HOT theories. If one part is responsible for making the mental state conscious, this logic implies that another part must make the first part conscious, and this would go ad infinitum. In response to this criticism, proponents might argue that it is sufficient for only certain parts of a mental state to be conscious in order for the entire state to be considered conscious, and this weaker criteria for consciousness are the solution to the worry of regress. However, it remains challenging to establish clear and

consistent criteria for determining precisely how many or which parts must be conscious for the complex state itself to count as conscious. Such a move works better if one argues for a modified HOT theory in which the complex becomes conscious through one part making another part conscious. Nonetheless, even this formulation lacks definitive criteria for favoring one form of higher-order intrinsic theory over another or preferring self-representationalism to higher-order intrinsic theories. Ultimately, HOT Theory prevents the regress by positing an unconscious HOT as responsible for consciousness. Critics who accuse HOT theory of employing an ad hoc strategy to circumvent infinite regress face similar challenges when proposing modified HOT theories. All relational theories confront the problem of regress because there is always the possibility of misrepresentation, as long as one state represents another.

Gennaro presents several reasons to favor intrinsicality. Firstly, as previously noted, there is an intuition that consciousness is intrinsic to mental states – a view supported both by common sense and phenomenological considerations. Rosenthal himself agrees that, even though consciousness need not mirror phenomenological considerations, a satisfactory theory of consciousness should do justice to our folk psychology: “Consciousness does reveal the phenomenological data that a theory of consciousness must do justice to. But to save these phenomena; we need only explain why things appear to consciousness as they do; we need not also suppose that these appearances are always accurate” (Rosenthal, 2004, p. 31). Kriegel also underscores the role phenomenological facts play in theories of consciousness. According to him, some phenomenological facts can lead us to infer some conclusions about consciousness. These facts are not justified with solid philosophical reasoning but still has a weight on our theory (Kriegel, 2009). Such a tendency is strongly present

in Galen Strawson's views and taken seriously by many philosophers in debates concerning the hard problem of consciousness. The problem of discussing intrinsicality in comparison to HOT theory is that Rosenthal does not sufficiently demonstrate that intrinsicality, as used in wide-intrinsicality view and self-representationalism, is in principle false.

Secondly, Gennaro supports WIV for its simplicity. With one complex mental state (CMS) it explains both the self-referentiality and world-directedness within the same structure. Of course, the practical difference between a single complex state with two components (WIV) and two separate extrinsic mental states (HOT theory) can be debated. There is a notable difference in the explanation of introspection. Rosenthal's explanation appears artificial in this respect, because it postulates another higher-order thought, which seems like an arbitrary extension of his primary explanation of world-directed consciousness (Gennaro, 1996, p. 27). On the other hand, self-referentialism is not in a better position regarding introspection, because self-referential structure is already integrated in world-directedness. Gennaro argues that WIV theory effectively synthesizes these theories and provides a better solution without resorting to ad hoc measures. Critics have challenged Rosenthal's reliance on a third-order thought to explain introspection, viewing it as an ad hoc response to problems like misrepresentation and infinite regress. In contrast, introspection in WIV occurs when a complex constituted of an unconscious meta-psychological thought (MET, designated M**) and a conscious MET (M*) is directed to the world-directed state (M). For Gennaro, this closed structure more naturally accounts for both consciousness and introspection without needing to expand the structure any further (with other extrinsic higher-order states). However, I do not think Gennaro's theory is evidently stronger than Rosenthal's open-ended

structure, wherein introducing a third-order thought legitimately explains introspection. Indeed, Rosenthal's framework is open to the possibility of additional higher-order states (such as a fourth-order thought and so on) without necessarily causing infinite regress – provided one does not insist that the conscious-making HOT itself must be conscious. Denying that consciousness is an intrinsic quality bolsters Rosenthal's position against regress issues. Nevertheless, if one insists that consciousness must be intrinsic and that the relationship between the world-directed state and the higher-order state should be non-contingent, WIV emerges as the stronger theory relative to HOT theory.

One might argue that intrinsic nature of consciousness does not preclude relational aspects, and consciousness need not be understood extrinsically to remain relational (Block, 2011). HOT theorists need stronger reasons for rejecting an intrinsic view of consciousness when addressing WIV and self-representationalism. At this junction, a decision between these theories hinge on other factors: Is it theoretically simpler and aesthetically preferable to have a unified complex state? Does such a structure carry an ontological advantage? Is it more effective to explain the self-referential element with a structure in which consciousness is intrinsic to the complex state? Gennaro responds positively to all these considerations, arguing that the phenomenological, intuitive and neurological evidence collectively support WIV (Gennaro, 2006, p. 224). He anticipates that the eventual discovery of the neural correlates of consciousness will likely reveal them as properties of a global conscious brain state.

Another problem WIV claims to address more effectively than HOT Theory is the "problem of the rock" (Stubenberg, 1998). This criticism aims to show that HOT theory either explains too little or too much. The question is: If one cannot

make a stone conscious by representing it with a HOT, how can one make a mental state conscious through a higher-order representation? A straightforward response is that HOT theory is a psychological theory, and only mental states have such a nature that they can become conscious through being represented by higher-order thoughts. However, the core of the criticism remains: Can HOT theory genuinely explain the hard problem of consciousness? Either it presupposes a kind of mentality that inherently includes consciousness, rendering the theory insufficient to explain phenomenal consciousness through higher-order thoughts, or it claims that higher-order thoughts alone fully explain our entire phenomenal realm, which appears excessively ambitious.

There is a significant relationship between the hard problem of consciousness and the problem of the rock. While the problem of the rock asks why objects like rocks do not become conscious when we direct our thoughts at them, the hard problem of consciousness asks how higher-order thoughts directed at mental states produce conscious experience. Thus, simply responding to the problem of the rock by asserting that rocks are not the right kinds of objects (mental states) does not adequately solve this problem (Kriegel, 2012, p. 75). Gennaro is on the right track when he criticizes Rosenthal's following answer to the problem of the rock:

> Being transitively conscious of a mental state does in a sense make it intransitively conscious. But that is not because being conscious of a mental state causes that state to have the property of being intransitively conscious; rather, it is because a mental state's being intransitively conscious simply consists in one's being transitively conscious of it. The mistake here is to suppose that a state's being intransitively conscious is an intrinsic property of that state. If it were, then being intransitively conscious could not consist in one's being transitively conscious of being in that state unless being thus conscious induced a change in that state's intrinsic properties. This objection is at bottom just a disguised version of the doctrine that being intransitively conscious is an intrinsic property. (Rosenthal, 1997, pp. 738-739)

Therefore, Rosenthal's theory should not be interpreted as a causal account of consciousness. Intransitive consciousness does not refer to an intrinsic property of a mental state but rather it is a term to facilitate explanations through transitive consciousness. A similar response can be given to Goldman's criticism of HOT theory, where Goldman questions how and why HOTs can confer consciousness on their objects (Gennaro, 2012, p. 72).

We should not insist that Rosenthal provides a causal explanation of consciousness. Self-Representationalism does not give a causal explanation either; instead, it articulates how self-consciousness is constitutive of the whole conscious state. However, we cannot entirely dismiss the problem of the rock, as it represents another aspect of the hard problem (Gennaro, 2012, p. 75) and HOT theory does not directly address the issue. While Rosenthal correctly notes that all these criticisms arise from the presumption that consciousness must be intrinsic to mentality and mental states, this does not invalidate the criticisms themselves. Highlighting that critics presuppose intrinsicality is insufficient to overcome their objections. Furthermore, arguing that intrinsic qualities result in non-informative and trivial explanations is not a strong or conclusive position.

3.4 Self-Representationalism

Self-representationalism attempts to integrate transitivity principle with phenomenological facts concerning intrinsic consciousness. In other words, it integrates the main idea of HOT theory, which bases consciousness on one's awareness of it (transitivity principle), with the phenomenological approach asserting that consciousness is essentially present in experience. The outcome is a relational

and representational account of consciousness in which consciousness remains constitutive and intrinsic to the whole mental state.

Kriegel argues that the phenomenal character is determined by the combination of mental qualities and subjectivity. Similar to Rosenthal, he presents different theories for qualitative and subjective character, instead of explaining consciousness and mentality under single account. Some form of representationalism is used to explain qualitative character of self-consciousness, while self-representationalism elucidates how the phenomenal character is constituted by the co-presence of qualitative and subjective character (Kriegel 2009, p. 53).

Kriegel's distinction between qualitative and subjective character comes from Levine (2001, pp. 6-7). Qualitative character refers to whatness of an experience: the bluishness of my visual experience of blue sky, the rectangularish quality of my experience of a box, the sweetness of my gustatory experience of a cheesecake are all examples of qualitative character (Kriegel, 2009, p. 45). It is not limited to perceptual experiences but extends to somatic and emotional experiences as well. Importantly, denying the existence of qualitative character is not on a par with denying the existence of qualia. Qualia are the properties of an experience which is "ineffable, intrinsic, private and directly apprehensible" (Kriegel, 2009, p. 47)[24] and Kriegel maintains a neutral position regarding qualia in this sense when distinguishing qualitative character from subjective character. Subjective character, in contrast, refers to the aspect of experience that makes it for-me. While the qualitative character determines what the experience is, subjective character makes it possible that there is an experience at all. Phenomenal character, then, is defined as the whatever property is responsible for the mystery of consciousness (Kriegel, 2009, p.

[24] We do not need to accept that a classical formulation of qualia necessarily includes direct apprehensibility. For our purposes, it is sufficient that qualia is ineffable, intrinsic and private.

53). Both qualitative and subjective characters contribute to discussions concerning the constitution of phenomenal character.

Kriegel introduces distinctions regarding subjective, qualitative and phenomenal characters to clarify his position and to highlight differences between various theories of consciousness. The first distinction is between *separatism* or i*nseparatism*. Separatism asserts that qualitative and subjective characters can be separated from one another, whereas inseparatism denies this. Notably, this separability need not be reciprocal: One could argue either (a) both qualitative and subjective characters are mutually separable, or (b) qualitative character is separable from subjective character, but not vice versa. Although choosing between (a) and (b) may not significantly impact practical discussions, these distinctions highlight intriguing metaphysical possibilities. For example, if both qualitative and subjective characters are mutually separable, then independent qualitative experiences could exist without a subjective character, and a free-floating subject could exist without a qualitative character. Alternatively, if qualitative experience is separable from subjective experience, but not vice versa, it implies that an experience with qualitative character could occur without a subject, while there could not be a subjective experience without qualities.

The second distinction is between (i) *subjectivism*, (ii) *qualitativism* and (iii) *compresentism.* Subjectivism claims that the subjective character determines phenomenal character. Qualitativism, in contrast, holds that qualitative character determines phenomenal character. For compresentism, phenomenal character emerges from the co-presence of qualitative and phenomenal character. First-order representationalist adopt *separatist qualitativism*: They first isolate qualitative character and give an account of mental qualities. Subsequently, they reduce

phenomenal character to qualitative character. HOT theorists typically argue for *separatist subjectivism*: They argue that subjective character, distinct and separable from qualitative character, ultimately determines phenomenal character. In this approach, different theories are given for qualitative and subjective characters but ultimately the higher-order thought overdetermines conscious experience. Self-representationalism addresses the overdetermination by HOT problem by proposing that conscious experience is a synthesis of subjective and qualitative characters.

Kriegel aligns with HOT theory and first-order representationalism in maintaining that qualitative character is separable from subjective character. However, he is inclined to defend *separatist compresentism*: He holds that qualitative character is separable from subjective character (without explicitly taking a stance on whether subjective character can likewise be separated from qualitative character), and he argues that the phenomenal character emerges from the co-presence of qualitative and subjective characters. Interestingly, Kriegel's position can also be interpreted as *inseparatist subjectivism*, wherein phenomenal character is determined by the subjective character, which itself is inseparable from qualitative character. It is not strictly necessary for his theory to conform exactly to a single position, but exploring these alternatives helps to clarify one's philosophical commitments and their broader implications. These distinctions are implicitly embedded in most theories of consciousness, even if not explicitly stated. According to Kriegel, a common issue across various theories of consciousness is that theorists often presume these distinctions without offering adequate justification or sufficient argumentative support.

There are various forms of self-representationalism, and unlike HOT theory, it lacks a single definitive version considered representative. Rather, it continues to

be a research program in progress. Despite its historical roots in the works of Aristotle and Brentano, self-representationalism continues to seek its core form. Uriah Kriegel, as the main proponent of this approach, presents different versions of self-representationalism, even supporting multiple versions within the same work. (Kriegel, 2009). However, one typical version defended by Kriegel (2006) can be considered as a well-rounded formulation of self-representationalism: "A mental state M of a subject S is conscious iff (i) M* is a (proper) part of M, (ii) M** is a (proper) part of M, (iii) M* represents M by representing M**, and (iv) M is a complex of M* and M**" (p. 153).

It is important to explain self-representation through the relation between two mental states while preserving it's intrinsicality. This is established by conceptualizing a mental state as a complex composed of two (proper) parts. In this way, consciousness is intrinsic to the complex mental state but remains relational due to the representational connection between its two parts. As M* and M** are proper parts of the M, it is also ensured that they have a necessary and constitutive relation. According to Kriegel, this contrasts with extrinsic HOT Theory, in which the relation between representing and represented mental states is contingent. All these features render self-representational theory a viable alternative to extrinsic HOT theory. Furthermore, this complex-part structure enables the explanation of self-representation without encountering the problems associated with accounts of direct self-reference. Kriegel (2007) provides a general argument to support self-representational theory:

> 1. A mental state M of subject S is conscious when, and only when, S is aware of M;
> 2. Awareness of X requires mental representation of X; therefore,
> 3. M is conscious, when, and only when, S has a mental state M*, such that M* represent M.

> 4. Either M* and M do not entertain a constitutive, non-contingent relation, or they do;
> 5. There are good reasons to deny that M* and M do have a contingent relation (denying extrinsic HOT Theory);
> 6. There are good reasons to think that it is the case that M* and M entertain a constitutive and non-contingent relation; therefore,
> 7. Plausibly, M is conscious when, and only when, M is self-representing. (pp. 51-52)

Self-representationalism accepts the transitivity principle (premise 1) and argues that such awareness is achieved through a representational relation – specifically through the relation of being represented (premise 2). Additionally, SR accepts that a mental state becomes conscious when it is represented by another mental state (premise 3). Thus far, it aligns with HOT Theory. The critical distinction arises in premise 4: Self-representational theories emphasize a constitutive, non-contingent relation between M* and M, differentiating themselves fundamentally from extrinsic HOT theory.

There are several reasons why Kriegel advocates for a constitutive relationship between M* and M. Firstly, a constitutive relation prevents the possibility of misrepresentation. HOT theorist needs to deny a truism to deal with the problem: “For any subject S and time t, there is something it is like to be S at t iff there is a mental state M, such that (i) S is in M at t and (ii) M is conscious at t” (Kriegel, 2009, p. 131). In cases of misrepresentation, the subject S is considered to be in a mental state by being represented by a higher-order state, despite the absence of an actual M. In Chapter 2, we discussed possible solutions to this problem: Firstly, HOT theorists can opt to explain a technical notion of consciousness by giving up on their ambitions to explain phenomenal consciousness. Secondly, they can deny that consciousness is a property of the mental state and argue it is rather a property of the subject. Thirdly, they can assert that in cases of misrepresentation, the intentional object of M* does not correspond to any real object, suggesting that the property of

consciousness is instantiated by the intended objects rather than actual objects (Kriegel, 2009, pp. 129-139). Self-representationalism aims to eliminate the possibility of radical misrepresentation altogether. Despite previous criticisms, this approach offers a stronger position for resolving the misrepresentation issue by denying the dissociation between first-order and second-order states within conscious states.

Secondly, having an extrinsic HOT renders consciousness an extrinsic property, which leads to epiphenomenalism, if we focus on the causal route from M* to M. However, HOT Theory does not present itself as a causal thesis and rather it is treated constitutively (Kriegel, 2009, pp. 143-144). This leads to the problem of the rock, or the generality problem discussed earlier. If HOT Theory were causal, it would explain how a HOT confers consciousness on an otherwise unconscious mental state. Yet, this raises the question why a HOT cannot similarly affect non-mental physical objects such as rocks.[25] The common response is that only mental states can be made conscious by higher-order thoughts. Following this move, one should either deny a causal interpretation or downplay the distinction between mental states and conscious mental states. Extrinsic construal makes consciousness causally inert, and these problems are better dealt with self-representationalism, which treats consciousness as an intrinsic property (Kriegel, 2009, pp. 144-145).

We have previously pointed out that self-representationalism is not a causal theory either. However, it more successfully demonstrates how consciousness is a constitutive feature of the whole mental state – intrinsic to it. One must approach HOT theory with greater caution in this respect, as considering an extrinsic feature as constitutive of a conscious mental state is challenging. Denying that HOT is the

[25] A more relevant formulation of the critique would be asking why we cannot make any kind of bodily process conscious by directing a HOT.

cause of consciousness further complicates the position of higher-order theorists. Self-representationalism holds the advantage here since intrinsicality more effectively supports the notion of consciousness being constitutive of the mental state. HOT theory continues to face difficulties in establishing consciousness of first-order mental states through an extrinsic HOTs. HOT theories should either strengthen the relation between first-order and higher-order states or focus on the unique nature of HOT itself, distinguishing it from ordinary thoughts. It is difficult for HOT theorist to defeat self-representationalism by choosing the former option, because arguing for a constitutive relation between two levels of mental states risks transforming any higher-order theory into a variant of self-representationalism. The latter approach, on the other hand, risks evolving the HOT theory into a form of first-order thought theory, thus jeopardizing relationality and raising further questions about the nature of the relationship between HOT and FOR.

Self-representationalists argue that their approach explains the immediacy of higher-order representation more effectively. While Rosenthal agrees that the representational relation should be non-inferential and immediate, self-representationalists emphasize that immediacy is best understood within a structure where consciousness is intrinsic to a complex mental state and the two parts of the representational relation are necessarily interrelated as well. In contrast, Rosenthal focuses on the appearance of immediacy, arguing that it is more significant that the subject remains unaware of the underlying causal processes. For Rosenthal, whenever a person has a higher-order thought, the resulting conscious mental state appears immediate. However, the ordinary experience itself does not fully determine the nature of the relevant phenomenon; the immediacy experienced in conscious states does not exclude the possibility of other representational mechanisms at work.

Rosenthal thus attempts to find a balance between doing justice to the phenomenology of immediacy and what the theory independently implies. It is one of our tasks to correctly represent the phenomenology of our experience, but we should differentiate this from theoretical implications masked as common sense or intuition. Rosenthal reminds us that science frequently refuted intuitions initially regarded as self-evident. On the other hand, Kriegel aims to explain actual immediacy, proposing that consciousness being intrinsic to the mental state provides the most coherent solution.

3.5 Conclusion

We discussed whether we need to modify HOT theory toward self-representationalism. Different self-representational approaches have their own distinct agendas and may not spefically address the issues problematic for HOT theory. Reflexive theories primarily aim to solve either (a) the problem of infinite regress or (b) the duplication of content. However, HOT theory already addresses these issues effectively: The notion of unconscious HOTs provides a straightforward solution to infinite regress, and the duplication of content problem pertains specifically to higher-order perception theories, not to higher-order thought theory itself.

Kriegel's self-representationalism, on the other hand, employs indirect reference and phenomenological self-awareness to explain how a conscious mental state is directed both outwardly and inwardly. Nonetheless, various self-representational theories exist within the literature, each utilizing different explanatory tools to account for self-representation. What they have in common is the assumption that consciousness must be an intrinsic property of the mental state –

an assumption requiring independent justification. Additionally, a significant motivation for theorists such as Kriegel and Gennaro is to prevent the possibility of radical misrepresentation, a concern that HOT theory does not address at all.

If the goal of placing the first-order state and higher-order state within a meta-state is solely to establish an ontology in which consciousness becomes an intrinsic property to the mental state, there is no compelling reason to modify HOT theory into an intrinsic higher-order theory. Similarly, if the objective is merely to establish a stronger connection between first-order state and higher-order thought to prevent radical misrepresentation, transforming HOT theory into self-representationalism remains unnecessary.

However, the discussions have demonstrated that HOT Theory, in its classical formulation, is not without issues. Several important questions remain unresolved: (1) How do we adequately explain the distinction between successful veridical perception and radical misrepresentation? What specific role does first-order representation play in conscious experience? (2) If HOT indeed overdetermines conscious experience, what kind of place has self within the content of HOT? Self-representationalism has not yet given entirely satisfactory answers to these questions so far but these concerns continue to pose challenges for HOT theory. Consequently, it becomes necessary to revisit the first-order state to thoroughly understand the nature of mental qualities and to reexamine the notion of self to clarify its connection to higher-order thought.

CHAPTER 4

MENTALITY AND SUBJECTIVITY

In Chapter 2, we saw that the debates surrounding the problem of misrepresentation largely depend on how we connect the division of phenomenal labor, which explains mental qualities and consciousness independently, and the transitivity principle, rooted in the intuitive idea that a conscious mental state is one in which we are aware of ourselves as being in. Within higher-order thought (HOT) theory, the tension between regarding both first-order and higher-order states as necessary conditions and viewing an appropriate HOT as a sufficient condition cannot easily be dissolved. One option is to accept the overdetermination of HOT and leave out the relational structure of HOT theory. A related alternative involves considering consciousness primarily as a property of the subject and transitivity principle as the subject's awareness of being in a mental state, irrespective of whether the corresponding first-order state genuinely exists or not.

In Chapter 3, we examined whether consciousness could be better explained through a form of self-representational theory, and if higher-order thought theory should be modified into an intrinsic form or fully transformed into self-representationalism. Our analysis revealed that these alternatives do not demonstrate a substantial explanatory advantage over classical HOT theory. Intrinsic theories are based on the foundational idea that consciousness is an intrinsic property of mental states, yet this assumption lacks sufficient independent justification beyond phenomenological arguments. Thus, we concluded that HOT theory could retain its core principles without significant modification. However, unresolved issues remain regarding the precise nature of the relation between first-order representation and

higher-order thought, alongside the absence of a developed theory addressing how the self is integrated into the contents of first-order and higher-order states.

In Chapter 4, I will aim to bridge these gaps by remaining within the higher-order framework while addressing identified issues. Firstly, we will examine Rosenthal's discussion of mental qualities, questioning whether it really makes sense to explain mental qualities independently from consciousness. In the second section, I will argue that a proper account of self is necessary to complement HOT theory effectively and relate phenomenological approaches to the higher-order framework.

4.1 The relational nature of mental qualities

Rosenthal aims to give a naturalistic explanation for mental qualities. Central to this approach is the idea that one should give an account of mental qualities that is independent from consciousness; otherwise, the explanation would be theoretically biased and uninformative. He provides reasons why we have the intuition that mental properties seem conscious and intrinsic, and illustrates how treating them as not necessarily conscious qualities can help us to deal with various problems in a better way. Rosenthal argues that mental qualities can exist without being consciously perceived. The dissociation refutes the commonly held belief that equates mental processes, particularly qualitative mental processes, with consciousness.

Many writers simply identify qualitative states with conscious qualitative states, often by using terms such as qualia or raw feel to denote specifically conscious qualities. However, this practice pre-empts any discussion of whether qualitative states can occur unconsciously and encourages the idea that the problem of explaining qualitative character is inseparable from the problem of explaining consciousness. (Rosenthal, 2005, Ch. 5)

According to Rosenthal, there is a qualitative aspect of pain and perceiving a color, because these are mental states with specific qualities regardless of conscious awareness. "Sensory qualities are properties of mental states, not properties of our consciousness of those states" (Rosenthal, 2005, p. 179). There is only what-it-is-like to experience pain when that mental state is consciously apprehended. We distinguish between different mental qualities according to their qualitative properties, which remain fundamentally unaltered regardless of whether the states are conscious or unconscious. Paradigm cases for such a view are blindsight, peripheral vision, subliminal perception and masked priming experiments (Rosenthal, 2005, pp. 108-109; p. 238). Numerous empirical cases illustrate the existence of mental states occurring outside our conscious awareness. According to Rosenthal, a mental state is not conscious if one is not aware of it. Thus, categorizing these states as subconscious would constitute a theoretical decision that requires independent justification.[26]

We are only conscious of our mental states when we form HOTs about those first-order states. Consequently, what-it-is-likeness only emerges when we are aware of those states. This contributes to the intuition that we can only distinguish among mental states based on our subjective experience of these states. However, just as physical properties exist and can be differentiated independently of conscious

[26] Opponents would reject the idea that empirical cases like blindsight support the existence of non-conscious mental states. Instead, they might interpret blindsight as merely involving non-conscious information processing. However, such interpretations – that these states are neither mental nor qualitative – are influenced by the assumption that mental qualities necessarily involve consciousness. Rosenthal counters this understanding by arguing that unconscious mental states indeed possess qualitative properties, provided they account for perceptual distinctions and can be differentiated from perceptible (physical) properties. If the perceptual differences observed in blindsight were "due to non-mental, subpersonal properties, those nonmental properties could presumably also explain conscious perceptual differences, so that the conscious perceptual properties would be explanatorily idle" (Rosenthal, 2022, p. 254). Such a conclusion is not desirable for the opponent. Therefore, it is more coherent to claim that we distinguish between perceptual states based on the same mental properties, regardless of whether these states are conscious or not, and thus regard these states as qualitative rather than subpersonal. I thank Lucas Thorpe for bringing this criticism to my attention.

awareness, mental properties can also be classified independently from our being conscious of them. Families of mental properties hold relations of resemblance and difference, and they are distinguished according to these relational categorizations: "Just as with mental and physical color, mental roundness and triangularity resemble and differ from each other in ways homomorphic to the similarities and differences that hold between physical roundness and triangularity" (Rosenthal, 2005, p. 140).

How are mental properties without what-it-is-likeness qualitative? Rosenthal notes that according to a commonsense understanding a property is considered qualitative as it is "essentially the way consciousness reveals it to be" (Rosenthal, 2005, p. 142). Many philosophers used to argue that physical objects also possess qualitative properties, typically exemplified by color properties. However, contemporary views often regard color properties as merely apparent rather than truly qualitative. "It is often concluded that physical objects do not literally have the properties of being red and green or, if they do have such properties, they are not as common sense conceives them" (Rosenthal, 2005, p. 160). Thinking otherwise reflects a judgment based on commonsense intuition that is not reliable. Similarly, Rosenthal argues that the commonsense intuition of attributing distinctive qualitative properties to sensory states is equally unreliable. He asserts, "we need not preserve the 'element of truth' in erroneous commonsense intuitions when we become convinced that these intuitions reflect only how things appear, rather than how they really are" (Rosenthal, 2005, p. 143). Thus, commonsense provides limited insight into sensory qualities beyond indicating that "they are those properties in virtue of which we distingush among those sensations, and that we can be more or less immediately conscious of them" (Rosenthal, 2005, p. 143). While mental qualities can be revealed only through consciousness, this does imply that they always occur

consciously. We can reveal their qualitative properties through forming HOTs, yet these properties are classified and distinguished by their mental qualities, independent of their conscious manifestation.

We might instead ask the question differently: Why do we assume that mental properties can only be considered qualitative when accompanied by what-it-is-likeness? This presumption arises from the idea that we fix mental qualities only through our consciousness of them. A primary motivation for this perspective stems from the effort to relocate qualitative character from physical objects into mental states. According to this view, as a physical object can be described mathematically, it cannot be qualitative. However, since it is thought that there must be something inherently qualitative about the color red, we attempt to "relocate the apparent qualitative character of physical color inside, into the mind" (Rosenthal, 2005, p. 142). A form of dualism underpins this approach: Physical objects themselves are not qualitative but give rise to visual sensations possessing qualitative properties. Rosenthal (2005) refers to this as *the relocation story* and notes Shoemaker's notion of *displacement maneuver* as an influence (p. 160).

In this picture, qualitative properties are thus viewed as intrinsically conscious, with the external world conceptualized as physical by transferring qualitative properties from the external, physical realm to the internal, mental realm. Further relocation is prevented by the claim that "there is no place to relocate the qualitative character that the distinctive properties of sensory states seem to exhibit" (Rosenthal, 2005, p. 142). However, Rosenthal argues that our inability to replicate this relocation does not compel us to conclude that mental properties must be qualitative in a sense that necessarily involves consciousness. Indeed, this is precisely what he seeks to establish through his distinct theories of mental qualities

and consciousness: "Qualitative properties are simply the distinguishing properties of qualitative states, the properties in virtue of which we classify them into mental kinds" (Rosenthal, 2005, p. 161).

Nothing in this notion inherently demands that qualitative properties must be conscious.[27] Although we might want to preserve certain properties that resist mathematical description, we can "characterize mental qualities by reference to commonsense perceptible properties" without reconstructing them with a built-in consciousness (Rosenthal, 2005, p. 12). Thus, a state's qualitative character can be explained through its perceptual role and relationally within a quality space, while higher-order thought theory separately accounts for how the subject becomes conscious of this qualitative character.

4.1.1 Quality space theory

If mental qualities are not intrinsically defined by consciousness, how are they individuated? Rosenthal defends a relational theory of mental qualities based on the concept of *quality space* structured by *just noticeable differences* (JNDs). A quality space is a multi-dimensional array representing the systematic similarities and differences among qualities within a particular sensory modality. For example, the color quality space encompasses dimensions such as hue, saturation, and brightness. Any specific color quality, such as a particular shade of red, is defined by its unique position within this space – a position determined by its relationship to all other color

[27] Depending on one's terminological preference regarding the concept of qualitative, the sentence could be alternatively stated as follows: "Nothing in this notion inherently demands that mental properties must be qualitative." If one wants to continue to equate qualitative properties with consciousness, one will have to maintain that mental states are not intrinsically qualitative but rather become qualitative only when conscious – an outcome achieved through their appropriate representation by a HOT. However, Rosenthal favors describing qualitative properties as not necessarily involving consciousness. Consequently, it is more appropriate within Rosenthal's terminology to say "nothing in this notion inherently demands that qualitative properties must be conscious."

qualities. Rosenthal (2005) emphasizes that "each sensory quality is the quality it is in virtue of the network of similarities and differences that obtain between it and all other qualities in the same family" (p. 161).

The structure of quality spaces is determined by the discriminative capacities of the organism. The metric of a quality space is defined by just noticeable differences (JNDs), which refers to the to the smallest (reliably) detectable difference between two stimuli by a subject.[28] Sensory states are considered qualitatively identical if there are no JNDs between them, and they differ qualitatively to the extent that JNDs separate them along relevant dimensions. The JNDs for any sensory modality determine a space of possible sensory qualities for that modality. Each quality is identified by its relative location in that quality space, which is determined by its JND relationships to all other qualities within the same space. Therefore, sensory qualities are fixed by these JND relationships (Rosenthal, 2005, p. 170).[29]

The distinguishing feature of a particular shade of red is not an ineffable, intrinsic property; rather, it is characterized by its unique relational profile, both

[28] One might challenge the usage of JNDs to explain mental qualities by arguing that reliable detection involves a behavioral measure. However, Rosenhal's reliance on JNDs does not imply a reduction of mental qualities to physical properties or behavioral dispositions. Instead, Rosenthal employs JNDs within his quality space theory as functional tools to systematically map and organize mental qualities. JNDs are tools to fix mental qualities in a relational and interrelated quality space. Thus, Rosenthal's aim is to provide an objective characterization of mental qualities and their comparative nature, rather than reducing mental qualities to behavioral responses. I would like to thank Lucas Thorpe for drawing my attention to this criticism.

[29] There might be cases in which individuals can reliably distinguish between the mental qualities yet report no apparent conscious difference. Such scenarios align closely with quality space theory and HOT theory of consciousness. According to HOT theory, only those mental states accompanied by appropriate HOTs become conscious. We constantly distinguish various mental qualities because we fix their references comparatively through just noticeable differences. These qualities are mental because they differ from physical/perceptible qualities, and theoretically, they are defined as properties situated within an interconnected mental space (quality space). Mental qualities are differentiated in the same way, whether conscious or unconscious. A subject would claim an apparent difference only when he was conscious of the corresponding mental states through appropriately formed higher-order thoughts. The content of the HOT determines the nature of experience, which can be more or less fine-grained compared to unconscious mental state. Consequently, a subject might report no conscious difference between two mental states if he lacks a sufficiently detailed conceptual repertoire relevant to the perception in question. For instance, one might unconsciously differentiate between the tastes of two different wines without being able to consciously identify or articulate this difference. I am indebted to Lucas Thorpe for highlighting this challenge.

inter-modally and intra-modally. This relational profile encompasses perceived similarities and contrasts with other specific colors, its relative position on the brightness scale, and related differentiating features. Ultimately, these relational attributes are defined by dispositions to differentiate. As Rosenthal (2005) notes, "The qualitative character of a sensory state, therefore, pertains to its dispositions to be discriminated from and analogized with other states" (p. 180).

A mental quality is defined by its relational comparisons to other qualities and is embedded within a complex system. Mental qualities are homomorphic to perceptible qualities of external objects (physical properties). Thus, the similarities and differences among mental qualities mirror those among perceptual properties. JNDs, at least in principle, help us to map the entire quality space of mental properties, even if this task remains practically unfeasible. Nonetheless, we can say that each mental quality is fixed by its position in this quality space and being conscious of that quality fixes the conscious qualitative character of that mental quality: “Each conscious qualitative state has the conscious qualitative character it does because one is conscious of that state as occupying a position in a mental quality space homomorphic to that of a particular property in the corresponding space of perceptible properties” (Rosenthal, 2005, p. 203).

4.1.2 Addressing the hard problem of consciousness

Rosenthal's relational theory of mental qualities provides a framework to deal with the hard problem of consciousness. The explanatory gap, as originally articulated by Levine (1983), refers to the difficulty in explaining how physical processes in the brain give rise to subjective qualitative experiences. Chalmers's (1995) formulation of the hard problem highlights a related concern, questioning why physical processes

are accompanied by conscious experience and why those experiences have the specific phenomenal character they do. Rosenthal's approach addresses these challenges by separating the problem of mental qualities from the problem of consciousness. If mental qualities are defined by their relational positions within a quality space, then explaining a particular mental quality involves explaining how the brain implements such relational structures and their associated discriminatory dispositions.

The what-it-is-like character is not an additional non-physical aspect; rather, it emerges from the subjective recognition of relational differences. When a mental state possessing a specific quality becomes conscious through a higher-order thought, the individual becomes aware of that quality's particular location within the relational quality space. Consequently, the subjective experiential quality arises from those discriminations that are either actively made or potentially available to be made. This approach reduces the explanatory gap by framing consciousness as a complex relational structure rather than a mysterious intrinsic property.

Just as our conscious awareness of physical properties does not necessarily reflect their real nature, our consciousness of mental qualities need not accurately represent their qualitative nature. Similarly, just as our perception of physical properties does not alter their physical nature, our being conscious of mental properties does not change their mental characteristics. As Rosenthal (2005) puts it, "the way we are conscious of the qualitative properties of sensory states when those states are conscious need not reflect the character of the mental qualities … mental qualities appear recalcitrant to mathematical description only because of the way we are conscious of those qualities when the relevant qualitative states are conscious" (p.

173). Nevertheless, mental properties retain the same nature and serve to differentiate mental states in the same manner, whether or not they are conscious.

The inverted spectrum thought experiment (Shoemaker, 1982; Block, 1990) posits two individuals whose color experiences are systematically inverted relative to each other despite identical behavioral responses and linguistic usage. This is commonly employed to argue that functional accounts cannot account for the intrinsic quality of experience. In such a case, JND relationships of the two individuals would be perfectly isomorphic – that is, their internal quality spaces would share identical structures of similarities and differences, despite differences in underlying neural realizations. However, according to Rosenthal's relational theory of sensory qualities, these individuals would still possess identical type*s* of color mental qualities: "two individuals whose color experiences were systematically inverted with respect to each other would nonetheless share all the same color qualities, provided the inversion preserved all JND relations" (Rosenthal, 2005, p. 186). Thus, as long as the homomorphic relations remain unchanged, their color experiences should also remain consistent.

The intuition that a change must have occurred relies on the idea that an intrinsic quality is inverted. However, if all relational properties are preserved, the mental qualities remain identical. Any purported difference that does not affect the structure of the quality space and discriminatory capacities is either non-existent as a mental property or irrelevant. The real challenge presented by the inverted spectrum presupposes the notion of intrinsic qualia that a relational theory aims to replace.

What would Mary learn within this homomorphic theory of mental qualities? Frank Jackson's (1982) thought experiment describes a neuroscientist, Mary, who learns everything about color vision while confined in a black-and-white room. Upon

leaving the room and seeing a red tomato for the first time, Mary seems to acquire new knowledge – specifically, the experiential knowledge of what it is like to see red. This suggests that physical facts alone cannot capture qualitative experiences.

Rosenthal argues that Mary does not learn a new non-physical fact but obtains new concepts or ways of representing color experiences. Before going out of her room, Mary could theoretically grasp the relational structure of the color quality space, including all just noticeable differences. When Mary sees red for the first time, she experiences a specific sensory state. By forming a higher-order thought about this state, she becomes consciously aware of it. What is new for Mary lies in her first-person experience of the quality and her ability to conceptualize it experientially.

Mary thus learns to apply her first-person concepts to mental states she already knew theoretically. The quality itself remains defined by its position in the quality space; what changes for Mary is her new mode of access and conscious experience of that quality. Although Mary gains a new experiential understanding in terms of what-it-is-likeness, she does not acquire new factual knowledge of a new mental property. The mental property itself remained the same before and after Mary was in the room, because it is defined relative to Mary's quality space of colors.

Jackson's argument against physicalism hinges on Mary acquiring new factual knowledge. According to Rosenthal (2019), the difficulty posed by his argument cannot be resolved either by adopting a consciousness-based approach, where consciousness is taken as intrinsic to the mental states, or employing Rosenthal's quality space theory, where mental states are determined relationally and independently of conscious awareness. Under the consciousness-based approach, Mary sees something new, but the knowledge she acquires is not factual. In contrast,

according to the quality space theory, Mary's seemingly new knowledge is factual but not actually new, as it was already present in Mary's books and available to her through her prior theoretical understanding (p. 60).

Differences in quality spaces affect both the types of mental qualities we have and the ways we are conscious of these mental states. Mary's quality space, for example, differs significantly from ours, as it exclusively includes comparative relations between shades of gray without being able to incorporate comparisons to other color qualities. Accordingly, Mary's new sensation upon seeing red will differ from our own experience of seeing red. The mental qualities are defined comparatively, "as resembling and differing from other visual states" (Rosenthal, 2005, p. 214). Thus, Mary's initial experience of seeing red would align with our own first-time perception of red only if one endorsed a consciousness-based view of mental qualities. In other words, "only if what it's like for one to be in a qualitative state were a matter of how one is conscious of the relevant quality on its own would the first conscious perception of something red be like our own" (Rosenthal, 2005, p. 214).

In contrast, mental states are fixed comparatively rather than through the way we are conscious of them. Thus, the what-it-is-like aspect of seeing red (for the first time) depends on the specific quality space within which the experience occurs. Consequently, an average person would be able to situate and identify the color red within a broader range of color comparisons than Mary could initially:

> Since the subjective color comparisons available to Mary are far fewer than ours, the descriptive content of her thought or remark, 'This is what it's like to see red,' will be different from the content of anything we might say or think. Mary's content will be roughly that the new experience is very different from anything she has ever seen, along with some comparisons with various gray-scale shades. By contrast, when we see something red . . . our descriptive content involves locating our experience among a dramatically greater range of colors. (Rosenthal, 2019, pp. 58-59)

In summary, there is something-it-is-like for Mary to see red for the first time, even though this experience is very different from what an ordinary person undergoes upon seeing red for the first time. According to Rosenthal, we can coherently state that Mary's experience is novel, but this does not constitute new factual knowledge. "If that knowledge is factual, it will be in Mary's book" (Rosenthal, 2019, p. 60). Consequently, this what-it-is-likeness does not cause any problem for physicalism.

Jackson construes knowledge of what-it's-like to see red as the subjective experience typically encountered by individuals when seeing red. If we reject Jackson's conception of factual knowledge, then Mary does not acquire new factual knowledge upon seeing red. Rosenthal concludes that Jackson's mistaken view that Mary gains new factual knowledge results from conflating two incompatible understandings of mental qualities (Rosenthal, 2019, p. 61).

4.1.3 Do Rosenthal's views on mental qualities help us to deal with the hard problem

We have seen that Rosenthal articulates a coherent position wherein he can distinguish clearly among physical properties, mental properties and consciousness. He effectively defends a relational account of mental properties, resisting hard problem of consciousness. However, ultimately, Rosenthal proposes a functionalist account of mental properties without explicitly demonstrating the possibility of reducing mental properties and consciousness. Nonetheless, his framework significantly challenges the intuition that consciousness must be intrinsic, monadic and non-relational. If Rosenthal successfully undermines the authority of this intuition, he successfully counters higher-order intrinsic theories and self-representationalism.

The intuition of intrinsicality stems partly from underestimating phenomenal richness afforded by a relational account and higher-order thoughts. A unique position in a high-dimensional quality space, defined by a complex web of JNDs, specifies a specific and determinate property. This characterization is neither thin nor impoverished. The what-it-is-like quality of a conscious experience thus emerges as the subjective apprehension of this specific and relationally defined location within the quality space.

Another source contributing to the richness of experience is the conceptual content of higher-order thoughts. Since the content of higher-order thought determines our conscious experience, these HOTs can help to have fine-grained experiences. Learning new words and acquiring new concepts enrich our perceptions. In this regard, both quality space theory and HOT theory collaboratively explain how we can have a rich mental life. The relational structure of our mental qualities enables us to be aware of endless possibilities of qualitative properties, whereas HOTs shape our subjective experience (what-it-is-likeness) in ways contingent upon our conceptual repertoire.

Moreover, a purely intrinsic quality devoid of relational and dispositional significance would be philosophically problematic. As Rosenthal (2005) asks: "If a property had no implications for dispositions to discriminate and recognize, it is unclear how it could count as a sensory quality at all. What would make it *that particular* quality, rather than another, or none at all?" (p. 190). Without relational properties, intrinsic qualities risk becoming explanatorily vacuous and epistemically inaccessible from a comparative standpoint.

4.2 First-order thought theory

Throughout the thesis, we have argued that we do not have to transform HOT theory into some form of higher-order intrinsic theory or self-representationalism. Nevertheless, an important challenge remained: If the content of higher-order thought overdetermines the nature of our conscious experience, what precisely is the exact role of first-order state?[30]

We examined first-order states and mental qualities, finding that Rosenthal's views on qualitative properties provides plausible answers to various problems. The key conclusions drawn are as follows: (1) The intuition of intrinsicality can be effectively explained away; (2) a relational account of mental properties can address the hard problem of consciousness and present a theoretically viable alternative; and (3) clearly distinguishing consciousness from mental properties offers significant theoretic advantages.

However, we can also recognize that intrinsicality is not the sole point of discussion between HOT theories and self-representational approaches. A crucial question remains: Are mental qualities without consciousness conceptually adequate? Additionally, a perhaps more pressing issue arises: How should we characterize the relation between first-order and higher-order states if HOT alone is responsible for the subjective what-it-is-like character of conscious experience?

One should either develop a HOT theory which necessitates the presence of both first-order and higher-order states or propose some variant of first-order thought theory. It is not merely an academic trend that philosophers persistently debate the problem of misrepresentation. Rather, the problem of misrepresentation serves as a crucial tool for examining the core nature of HOT theory – particularly its approach

[30] You can find a discussion regarding the causal role of first-order states in Chapter 2.

of explaining phenomenal consciousness solely through HOT, despite traditionally emphasizing the relation between first-order and higher-order states.

4.2.1 Non-relational HOT theories

Our analysis of HOT theory through the problem of misrepresentation revealed that the theory oscillates between employing the transitivity principle and utilizing the phenomenal division of labor. The phenomenal division of labor necessitates assigning distinct explanatory roles to first-order and higher-order states, whereas the transitivity principle ties conscious experience directly to the higher-order thought. Importantly, according to this principle, the subject need not be in the first-order mental state at all; rather, conscious experience is primarily determined by the content articulated by the higher-order thought.

Brown introduces several distinctions which can complement this analysis of HOT theories. First, he distinguishes between *traditional* HOT theories, which accept the transitivity principle (TP), from *non-traditional* HOT theories, which either reject or reinterpret TP. Brown denies TP in its classical formulation – the claim that "a conscious state is a state whose subject is, in some suitable way, aware of being in it" – and situates himself within the non-traditional camp (Brown, 2025, p. 145). According to Brown, all theories examined in this thesis belong to the category of traditional HOT theories. Traditional HOT theories explicitly commit to the state-consciousness of the first-order state, asserting that "it is the first-order state itself that is like something for one to be in, and the explanation of that is given in terms of the transitivity principle" (Brown, 2025, Ch. 1). Brown also draws a conceptual distinction between *phenomenal consciousness* – "the state which, when I am in it, there is something that it is like for me to be in it" – and *state consciousness*

– "the state of which I am aware" – (Brown, 2025, pp. 71-72). He argues that in ordinary conscious experience, the first-order state is state-conscious, but not phenomenally conscious, while higher-order state is phenomenally conscious but not state conscious.

The first-order state is state-conscious, because one is aware of the state targeted by the higher-order state. As Brown (2025) explains, "the things we count as being aware of are the states that are referred to, or picked out by, the proposed targeting relation" (p. 72). However, first-order states cannot be considered as phenomenally conscious, as they are not *phenomenologically manifest* – where phenomenological manifestness describes "the way in which we are aware of our own mental life"[31] (Brown, 2025, p. 73). According to this characterization, higher-order states can be classified as phenomenally conscious states without being state-conscious. Brown thereby rejects the transitivity principle by highlighting the possibility of conscious states of which we are not aware, a scenario in direct conflict with the classical formulation of TP.

A crucial distinction emerges at this point. *Relational HOT* theories hold that the content of the first-order state directly contributes to phenomenal consciousness. In contrast, *non-relational HOT* theories argue that "first-order state does not directly affect phenomenology. Instead, higher-order representation alone determines what it's like for the subject" (Brown, 2025, p. 145). Under this characterization, Rosenthal's classical formulation of HOT theory, Gennaro's wide intrinsicality view and Kriegel's self-representationalism all fall within traditional non-relational HOT

[31] Brown defines the notion of phenomenologically manifest as follows: "Something is phenomenologically manifest when it is conveyed to the subject by one's phenomenology. To be phenomenologically manifest is to be occurrent in one's stream of consciousness … phenomenal character is what is overall phenomenologically manifest, and the specific things that are manifest in this way are the phenomenal properties of one's phenomenally conscious states" (Brown, 2025, p. 28).

theories. The distinctions among these theories lie in subtle details: Rosenthal represents *pure non-relationalism,* as he does not secure the existence of first-order state. Conversely, Gennaro and Kriegel, who attempts to secure the first-order representation, can be regarded as *mixed non-relationalists.* However, because they ensure the presence of the first-order state only through a form of indirect reference without attributing a decisive role to its content, they do not qualify as relational HOT theorists.[32]

Brown (2025) argues that the traditional HOT theorists accept two central premises: First, TP defines "the data a theory of consciousness needs to account for", and second, that phenomenal consciousness is explicable through the state-consciousness of a first-order state, meaning there is something-it is-like to be in that first-order state (p. 144). Accordingly, disagreements among traditional HOT theories stem from an ambiguity in the interpretation of the transitivity principle, leading to relational and non-relational versions of HOT theories. Relational HOT theories hold that consciousness is "a property of the first-order state that is a result of one's being in an awareness relation to that very state" (Brown, 2025, p. 418). This approach conceptualizes consciousness explicitly as a relation between first-order and higher-order states. It explains consciousness of the first-order state and affirms that "what it is like for the subject is to some degree determined by the first-

[32] Rosenthal thinks that in the case I think about an object x, the relation between me and the intentional object of my thought is not genuine. The object does not participate in my conscious thought of x and need to exist. On the other hand, if I am aware of an object y, a real relation appears to hold between my awareness and y, because "one cannot be aware of something unless that thing exists" (David Rosenthal, personal communication, June 3, 2025). In this case, the object y refers to myself*;* thus, I am aware of myself as being in a particular state. Considering these, Rosenthal finds the relational vs non-relational categorization not so helpful – considering no traditional higher-order theory is relational in Brown's sense – and somewhat confused (David Rosenthal, personal communication, June 3, 2025). While Brown focuses on the relation between first-order and higher-order states, Rosenthal emphasizes the distinction between *me* and *myself*. According to my interpretation, Rosenthal would consider the former relation to be non-genuine, whereas the latter one a real one. I am grateful to David Rosenthal for his feedback on this issue.

order state's content (the first-order state is phenomenally active)" (Brown, 2025, p. 148). Conversely, non-relational HOT theories accept the state-consciousness of the first-order state as the explicandum but deny that "the first-order state targeted by the higher-order representation determines what it is like for the subject … The first-order state of which we are aware will be phenomenally silent" (Brown, 2025, p. 149).

Brown presents his theory as a form of non-traditional HOT theory mainly through two aspects. First, he emphasizes inner awareness as a form of higher-order representation, rather than a perception or a thought. Traditionally, "there are two ways we can be conscious of things: By perceiving them and by having thoughts about them" (Rosenthal, 2005, p. 145). Brown argues that this understanding is based on folk-psychology and relying on cognitive sciences lead us to understand inner-awareness as a form of (higher-order) representation. It is higher-order in the sense that it represents something mental, and it is a representation, meaning it "consists in a state that places satisfaction conditions on the thing being represented" (Brown, 2025, p. 66). Second, it accepts a revised formulation of transitivity principle (TP*) : "a mental state is phenomenally conscious if and only if it appropriately makes one aware of one's mind" (Brown, 2025, p. 72). This formulation of transitivity principle captures phenomenal consciousness and not state consciousness, because state consciousness "applies to the state that one is aware of, not the state of awareness itself" (Brown, 2025, p. 72).

4.2.2 Qualitative character and subjective character

I will not engage here in the discussion of whether Brown makes the correct distinctions; nevertheless, his categorization and arguments supporting this mapping

are intriguing. Brown makes use of Kriegel's (2009) distinction between qualitative and subjective character. Accordingly, qualitative character determines the whatness of an experience, while subjective character determines that an experience is for-me. Together, qualitative and subjective characters constitute phenomenal character, which is the complete what-it-is-likeness of an experience for the subject.

Brown interprets the distinction in the following way: Kriegel's concept of qualitative character resembles Rosenthal's notion of mental qualities, although Kriegel's qualitative character involves consciousness to a greater extent than Rosenthal's mental qualities. On the other hand, Kriegel's notion of subjective character aligns more closely with Rosenthal's conception of consciousness. Within this framework, HOT theory can be understood as primarily emphasizing subjective character, which thereby overdetermines conscious experience. According to Brown, a reinterpretation of Rosenthal's account of mental qualities in a manner that aligns more closely with Kriegel's qualitative character could render mental qualities in HOT theory genuinely phenomenological. The challenge, however, lies in the fact that Kriegel's distinction between qualitative and subjective character assumes an intrinsic dimension of consciousness that is not present in Rosenthal's relational account of mental qualities. Nevertheless, Kriegel's proposed subjective dimension might prove beneficial in developing an enhanced version of HOT theory.

Thus, Brown attempts to apply the notions of qualitative and subjective characters to reinterpret Rosenthal's concepts of mental quality and consciousness. The first step involves employing a form of recognitional concept, enabling existing mental qualities to be referred to by higher-order thoughts and thereby acquiring subjective character. Rosenthal explicitly rejects interpreting the concepts used in HOTs as recognitional or phenomenal concepts (Rosenthal, 2005, p. 207). However,

he utilizes a very thin notion of the self, and this allows for integrating the notion of subjective character into HOT theory with minimal theoretical modifications

In this context, we can analyze the debate between self-representational accounts and Rosenthal's HOT theory through the lens of the priority of self-consciousness. Kriegel (2004) argues that self-consciousness is prior to consciousness, particularly emphasizing that *intransitive state self-consciousness* precedes all other forms of consciousness. This priority reflects a dependency relationship influenced by phenomenological approaches, according to which a minimal, pre-reflective form of self-consciousness is inherently embedded within our experiences. Zahavi (2005) argues that phenomenal consciousness cannot be explained through a higher-order representation, and it should be rather investigated as the intrinsic, pre-reflective subjectivity that makes an experience for-me. Clearly, this position directly contrasts with Rosenthal's views. The pertinent question, therefore, is whether insights from these phenomenological and self-representational perspectives could fruitfully inform the development of a revised version of HOT theory.

We can be optimistic in this regard. Kriegel is significantly influenced by Rosenthal's terminology, particularly Rosenthal's distinctions between state consciousness and creature consciousness, as well as transitive and intransitive consciousness. He constructed his self-representationalism by integrating elements from HOT theory with phenomenological approaches. Brown goes a step further by categorizing both HOT theory and self-representationalism as non-relational HOT theories. Given these convergences, how can we effectively incorporate a dimension of subjectivity into HOT theory?

The tools Kriegel uses in this direction do not appear promising for HOT theory. Kriegel particularly relies on the notion of peripheral awareness to account for self-consciousness that resides in the background of our experience. Additionally, he utilizes indirect reference to provide a naturalistic explanation of self-reference. Both Gennaro and Rosenthal find these methodological tools problematic.

However, Brown's emphasis on transitivity principle is instructive. The most coherent interpretation of a non-relational HOT theory involves understanding the theory as describing a relationship between a subject and the subject's mental properties. This interpretation yields a more robust formulation of transitivity principle. If TP is considered solely as a relation between two mental states, considerable effort must be devoted to demonstrating how a first-order state can simultaneously be present yet have no substantial impact on conscious experience. Conversely, interpreting TP as a relationship between the subject and their properties allows one to argue that the subject always has some mental states of which he is aware. Under this interpretation, cases of radical misrepresentation can be reclassified as standard instances of misrepresentation. It is also worth recalling that this approach to consciousness as a property of the subject provided a comparatively straightforward resolution to the misrepresentation problem discussed in Chapter 2.

There is also a gap in Rosenthal's views that requires further elaboration. While he has extensively addressed consciousness and mental qualities, his view on the self is limited. He defends a thin notion of the self as an embedded component within the intentional content of HOT. This conception of the self is minimal, referential and functional. Rosenthal explicitly rejects the notion of a substantive self as a metaphysical entity, as well as any phenomenal self directly present in first-

order states. Rather, he advocates a dispositional notion of self, invoked within the content of HOT when necessary:

> One's sense of self may be mainly dispositional, consisting in the various ways one would be disposed to think about oneself in various contexts. A purely dispositional sense of self contrasts with a self that is experientially or phenomenologically immediate and robust. (Rosenthal, 2005, p. 372)

> There need not be any actual, occurrent, phenomenally immediate self-awareness for a creature to have a sense of self. It is enough if the creature is disposed to think of itself, and this disposition gives rise to self-awareness when the right conditions occur. (Rosenthal, 2005, p. 375)

Introducing an occurrent notion of self may prove helpful and could integrate more naturally with the HOT framework. The classical conception of HOT theory, wherein higher-order thoughts can become conscious through third-order thoughts, appears overly structured. Within this classical approach, introspective consciousness is seen as an extension of the relational structure, a perspective whose difficulties have been addressed earlier in this thesis. Given that the content of HOT determines what-it-is-likeness, an occurrent notion of the self might be advantageous. Originally, Kriegel explains subjective character as being intrinsic to our mental states, while Rosenthal integrates it into the HOT theory as a dispositional feature. Considering Rosenthal's explicit opposition to usage of dispositional HOTs, an occurrent conception of the self might align well with the occurrent approach to HOTs.

Brown presents an intriguing argument against considering transitivity principle as a folk platitude, proposing instead that inner awareness should occupy this role. I consider this a compelling approach. HOT theories provide a non-relational explanation of consciousness through the content of HOTs, with subjective character playing a pivotal role. Grounding HOT theory in the notion of inner awareness offers a more robust explanation of HOT theory compared to analyzing it as fluctuating between transitivity principle and relational structures. Furthermore,

this perspective fosters a stronger alignment with self-representationalism and could establish unexpected alliances with phenomenological approaches.

This attitude also transforms a potential problem into an opportunity. Previously, we identified the risk that HOT theory could inadvertently become a first-order thought theory due to its tendency to higher-order thoughts overdetermining conscious experience, as highlighted in discussions of misrepresentation. The notion of free-floating thoughts that render notional states conscious presents significant difficulties.[33] However, interpreting the transitivity principle as the awareness relationship between the subject and their mental states considerably simplifies the theoretical task.[34]

[33] A HOT theorist would object to employing such a characterization. This objection could indeed be justified, as Rosenthal finds the causal role of first-order states in generating higher-order thoughts important. Nevertheless, he acknowledges that considerations of causality are ultimately irrelevant for the HOT theory of consciousness. There does not have to be causal ties between the intentional objects and higher-order thoughts. Consequently, the theoretical possibility of having free-floating higher-order thoughts without corresponding first-order states remains. You can check Chapter 2 for a discussion on causal roles of first-order states.

[34] In a certain sense, the risk of HOT theory shifting toward some variant of first-order thought theory remains. Brown explicitly states that his higher-order representational theory of consciousness is not a form of self-representationalism and, in fact, shares certain similarities with first-order representationalism. This similarity arises because his formulation of the transitivity principle (TP*) explains consciousness via an appropriate representation that "makes one aware of oneself as being in some first-order mental state" (Onemorebrown, 2021). He views this formulation as preserving beneficial aspects of first-order representationalism. Nonetheless, because Brown emphasizes inner-awareness in his explanation of consciousness, his theory is classified primarily as a higher-order theory according to his own categorization.

CONCLUSION

I used the misrepresentation problem as a methodological tool to explore central characteristics of HOT theory. While the possibility of misrepresentation is not inherently problematic, the way HOT theory deals with radical misrepresentation leads to multiple distinct formulations of the HOT theory.

I defended the view that we do not really need to modify HOT theory into a form of self-representationalism. However, I also argued that some of the challenges are not easily resolvable, underscoring the ongoing significance of the misrepresentation problem in philosophical discourse. There is a tension between the transitivity principle and the division of phenomenal labor. On the one side, higher-order thought overdetermines the nature of conscious experience, while on the other, HOT theory posits a relational structure between first-order and higher-order states.

I argued that self-representational approaches are motivated not only by the intuition that consciousness is intrinsic, but also the desire to secure the existence of the first-order state without rejecting the transitivity principle. Rosenthal's views on mental qualities bolster classical HOT theory, yet they do not conclusively favor HOT theory against alternatives theories.

I concluded by offering suggestions and establishing conceptual connections. Brown's distinction between relational and non-relational HOT theories is insightful and HOT theory can be described as a traditional non-relational theory. The transitivity principle should receive further critical examination within HOT literature, rather than being accepted as an intuitive common-sense starting point. Instead, inner awareness might provide a more robust foundation for higher-order thought theories.

Moreover, HOT theory requires a developed account of the self, beyond a dispositional or minimal understanding. In this context, additional comparative analyses involving the works of Kriegel, Zahavi and Rosenthal can be helpful.

REFERENCES

Aristotle & Shields, C. J. (2016). *De anima* (L. Judson, Ed.; C. J. Shields, Trans.). Clarendon Press Oxford University Press.

Berger, J. (2014). Consciousness is not a property of states: A reply to Wilberg. *Philosophical Psychology*, *27*(6), 829–842. https://doi.org/10.1080/09515089.2013.771241

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227-287.

Block, N. (2009). Comparing the major theories of consciousness. In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences* (4th ed.). The MIT Press. https://doi.org/10.7551/mitpress/8029.003.0099

Block, N. (2011). The higher order approach to consciousness is defunct. *Analysis*, *71*(3), 419–431.

Brentano, F. (1973). *Psychologie vom empirischen standpunkt*. Felix Meiner Verlag. (Original work published 1874)

Brentano, F. (1995). *Psychology from an empirical standpoint* (L. L. McAlister, C. Rancurello, & D. B. Terrell, Trans.; 2nd ed). Taylor and Francis. (Original work published 1874)

Brown, R. (2025). *Consciousness as representing one's mind: The higher-order approach to consciousness explained*. Oxford University Press.

Byrne, A. (1997). Some like it HOT: Consciousness and higher-order thoughts. *Philosophical Studies*, *86*(2), 103–129. https://doi.org/10.1023/A:1017959107565

Carruthers, P., & Gennaro, R. (2020, September 2). Higher-order theories of consciousness. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/consciousness-higher/

Caston, V. (2002). Aristotle on consciousness. *Mind*, *111*(444), 751–815. https://doi.org/10.1093/mind/111.444.751

Chalmers, D. (1997). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Coleman, S. (2015). Quotational higher-order thought theory. *Philosophical Studies*, *172*(10), 2705–2733. https://doi.org/10.1007/s11098-015-0441-1

Coleman, S. (2018). The merits of higher-order thought theories. *Trans/Form/Ação*, *41*(spe), 31–48. https://doi.org/10.1590/0101-3173.2018.v41esp.04.p31

Farrell, J. (2017). Higher-order theories of consciousness and what-it-is-like-ness. *Philosophical Studies*, 1–19. https://doi.org/10.1007/s11098-017-0980-8

Gallagher, S., & Zahavi, D. (2020). *The phenomenological mind* (3rd ed.). Routledge. https://doi.org/10.4324/9780429319792

Gennaro, R. (1996). *Consciousness and self-consciousness*. John Benjamins.

Gennaro, R. J. (2004). Higher-order thoughts, animal consciousness and misrepresentation: A reply to Carruthers and Levine. In R. J. Gennaro (Ed.), *Higher-order theories of consciousness: An anthology* (pp. 45–66). John Benjamins.

Gennaro, R. (2006). Between pure self-referentialism and the extrinsic HOT theory of consciousness. In U. Kriegel & K. Williford (Eds.), *Self-representational approaches to consciousness*. The MIT Press. https://doi.org/10.7551/mitpress/6155.003.0012

Gennaro, R. J. (2012). *The consciousness paradox: Consciousness, concepts, and higher-order thoughts*. MIT Press.

Gois, I. (2010). A dilemma for higher-order theories of consciousness. *Philosophia*, *38*(1), 143–156. https://doi.org/10.1007/s11406-009-9190-8

Gottlieb, J. (2020). On ambitious higher-order theories of consciousness. *Philosophical Psychology*, *33*(3), 421–441. https://doi.org/10.1080/09515089.2020.1731445

Grimes, J. (1996). On the failure to detect changes in scenes across saccades. In K. Akins & K. Akins (Eds.), *Perception*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195084627.001.0001

Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, *32*(127), 127–136. https://doi.org/10.2307/2960077

Janzen, G. (2011). In defense of the what-it-is-likeness of experience. *The Southern Journal of Philosophy*, *49*(3), 271–293. https://doi.org/10.1111/j.2041-6962.2011.00074.x

Kriegel, U. (2003). Consciousness, higher-order content, and the individuation of vehicles. *Synthese*, *134*, 477–504.

Kriegel, U. (2003b). Consciousness as intransitive self-consciousness: Two views and an argument. *Canadian Journal of Philosophy*, *33*(1), 103–132. https://doi.org/10.1080/00455091.2003.10716537

Kriegel, U. (2004). Consciousness and self-consciousness. *The Monist*, *87*(2), 182–205. https://doi.org/10.5840/monist20048725

Kriegel, U. (2006). The same-order monitoring theory of consciousness. In U. Kriegel & K. Williford (Eds.), *Self-representational approaches to consciousness* (pp. 143–170). The MIT Press. https://doi.org/10.7551/mitpress/6155.003.0008

Kriegel, U. (2007). Philosophical theories of consciousness: Contemporary western perspectives. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *Cambridge handbook of consciousness* (pp. 35–66). Cambridge University Press.

Kriegel, U. (2009). *Subjective consciousness: A self-representational theory*. Oxford University Press.

Kriegel, U., & Williford, K. (2006). Introduction. In U. Kriegel & K. Williford (Eds.), *Self-representational approaches to consciousness*. MIT Press.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, *64*(4), 354–361. https://doi.org/10.1111/j.1468-0114.1983.tb00207.x

Levine, J. (2001). *Purple haze: The puzzle of consciousness*. Oxford University Press.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914

Natsoulas, T. (1993). Consciousness$_4$: Varieties of intrinsic theory. *The Journal of Mind and Behavior*, *14*(2), 107–132.

Neander, K. (1998). The division of phenomenal labor: A problem for representational theories of consciousness. *Noûs*, *32*(S12), 411–434. https://doi.org/10.1111/0029-4624.32.s12.18

Onemorebrown. (2021, April 26). *Richard Brown -Higher-Order Theories of Phenomenal Consciousness* [Video]. Youtube. https://www.youtube.com/watch?v=6WZoPYYIOis

Picciuto, V. (2017). Keeping it real: Intentional inexistents, fineness-of-grain, and the dilemma for extrinsic higher-order representational theories. *Pacific Philosophical Quarterly*, *98*(4), 555–575. https://doi.org/10.1111/papq.12121

Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies*, *49*(3), 329–359.

Rosenthal, D. M. (1990). A theory of consciousness. In N. Block, G. Güzeldere, & O. Flanagan (Eds.), *The nature of consciousness: Philosophical debates* (pp. 729–753). MIT Press.

Rosenthal, D. M. (1993). Higher-order thoughts and the appendage theory of consciousness. *Philosophical Psychology*, *6*(2), 155–166. https://doi.org/10.1080/09515089308573085

Rosenthal, D. (1997). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates*. MIT Press.

Rosenthal, D. (2002). Explaining consciousness. In *Philosophy of mind: Classical and contemporary readings* (pp. 109–131). Oxford University Press.

Rosenthal, D. M. (2004). Varieties of higher-order thought theory. In R. J. Gennaro (Ed.), *Higher-order theories of consciousness. An Anthology* (pp. 17–44). Amsterdam: Johns Benjamins Publishing Company.

Rosenthal, D. M. (2005). *Consciousness and mind*. Oxford University Press. https://doi.org/10.1093/oso/9780198236979.001.0001

Rosenthal, D. M. (2005b). Consciousness, interpretation and higher-order thought. In P. Giampieri-Deutsch (Ed.), *Veröffentlichen der Kommission für Philosophie und Pedagogik* (pp. 119–142). Verlag der Österreichischen Akademie der Wissenschaften. https://doi.org/10.1553/3-7001-3386-3s119

Rosenthal, D. (2009). Higher-order theories of consciousness. In A. Beckermann, B. P. McLaughlin, & S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 239–252). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199262618.001.0001

Rosenthal, D. M. (2010). How to think about mental qualities. *Philosophical Issues*, *20*(1), 368–393.

Rosenthal, D. (2011). Exaggerated reports: Reply to Block. *Analysis*, *71*(3), 431–437. https://doi.org/10.1093/analys/anr039

Rosenthal, D. (2018). Misrepresentation and mental appearance. *Trans/Form/Ação*, *41*(spe), 49–74. https://doi.org/10.1590/0101-3173.2018.v41esp.05.p49

Rosenthal, D. (2019). There's nothing about Mary. In S. Coleman (Ed.), *The knowledge argument* (1st ed., pp. 32–61). Cambridge University Press. https://doi.org/10.1017/9781316494134.003

Smith, D. W. (1986). The structure of (self-) consciousness. *Topoi*, *5*(2), 149–156. https://doi.org/10.1007/BF00139228

Smith, D. W. (1989). *The circle of acquaintance*. Springer Netherlands. https://doi.org/10.1007/978-94-009-0961-8

Stubenberg, L. (1998). *Consciousness and qualia* (Vol. 5). John Benjamins Publishing Company.

Textor, M. (2017). Brentano on consciousness. In U. Kriegel (Ed.), *Routledge handbook of Franz Brentano and the Brentano School* (pp. 49–60). Routledge.

Weisberg, J. (2011). Misrepresenting consciousness. *Philosophical Studies*, *154*(3), 409–433. https://doi.org/10.1007/s11098-010-9567-3

Weisberg, J. (2014). *Consciousness*. Polity Press.

Weisberg, J. (2011b). Abusing the notion of what-it's-like-ness: A response to Block. *Analysis*, *71*(3), 438–443. https://doi.org/10.1093/analys/anr040

Wilberg, J. (2010). Consciousness and false HOTs. *Philosophical Psychology*, *23*(5), 617–638. https://doi.org/10.1080/09515089.2010.514567

Zahavi, D. (2004). Back to Brentano. *Journal of Consciousness Studies*, *11*(10–11), 66–87.

Zahavi, D. (2020). *Self-awareness and alterity: A phenomenological investigation*. Northwestern University Press.