Uniwersytet Warszawski

Szkoła Doktorska Nauk Humanistycznych

Maciej Tarnowski

Moore's Paradox in Language and Thought

Towards a Unified Strategy of Explanation

PhD thesis in philosophy written under the supervision of prof. dr hab. Tadeusz Ciecierski

Warszawa, 2024

Contents:

Figures:

Tables:

Acknowledgements

This work, as is often the case in philosophy, only disguises itself as a one-man achievement. People frequently tend to think of this branch of humanities as particularly detached from any need for collaboration. If that was ever a general truth, we are luckily past that time, and this dissertation is no exception to this trend. Only with immense support from others, this thesis could be completed: and all of them should be thanked here.

First thanks should be directed towards my supervisor, Tadeusz Ciecierski. After working with Tadeusz for nine years, I don't think I can sum up everything I learned from him in one or two simple sentences. Let me just say that I know no other person, who so effortlessly combines friendly support with insightful criticism, which, in my case, oftentimes led me to rethink everything from scratch, but always resulted in a substantial improvement. Thank you, Tadeusz, for all your help and advice.

But as it takes a village to raise a child, usually it takes a whole department to raise a graduate student. In this particular case, three departments were needed: my home Faculty of Philosophy at the University of Warsaw, the Department of Philosophy and Education Sciences at the University of Turin, where I spent the spring semester of 2023 thanks to Erasmus+ Program, and the Department of Linguistics and Philosophy at Massachusetts Institute of Technology, where I spent the fall semester of 2023 thanks to the Polish-American Fulbright Commission. My deepest thanks go to people who made me feel at home at those places: Joanna Odrowąż-Sypniewska, the chair of the analytic philosophy lab in Warsaw; Jan Sprenger, who supervised my work in Turin; Alex Byrne, my faculty sponsor at MIT; and all people who were a part of these vibrant philosophical communities at the time. It was a privilege to meet and work with you.

The previous incarnations of the material present in this dissertation, sometimes extremely distant from their present state, were presented at many different workshops and conferences. Those that led me to substantially change and improve on their content are Analytic Philosophy Lab seminars at the University of Warsaw, MIT's MATTI seminar, 16th Graduate Conference on Philosophy of Mathematics and Logic at the University of Cambridge, COGITO Research Seminar at the University of Bologna and FINO Philosophy Seminar at the University of Genoa; countless other events motivated me to organize my thoughts in a coherent manner and provided opportunity for clarifying and improving them, as well as gathering peer feedback. I hereby thank the organizers and participants of all these events. I am also grateful to National Science Center (NCN), who supported my work on the PRELUDIUM 20 grant (no. 2021/41/N/HS1/01586) which allowed me to attend these conferences and work on the contents of this dissertation throughout the last three years.

For reading and commenting on the drafts of the papers on which chapters 2-5 are based, I would like to thank Grzegorz Gaszczyk and Joanna Odrowąż-Sypniewska (Chapter 2.), Alex Byrne, Maciej Głowacki, Jan Sprenger and Juan Murillo Vargas (Chapter 4.), and Alex Byrne and Yonathan Fiat (Chapter 5.). Chapter 3. greatly benefited from the conversations with Joanna Odrowąż-Sypniewska, Grzegorz Gaszczyk, Juan Murillo Vargas, Matt Mandelkern, and Dilip Ninan. Parts of Chapter 1. also benefitted from a historical query in Cambridge University Library, where I got help and advice from Michael Potter. For commenting on the whole manuscript and suggesting various improvements, I am also (again) incredibly indebted to my advisor.

Though it's sometimes hard to pinpoint the exact point of influence, I want to also thank Krzysztof Sękowski, Adrian Ziółkowski, and Maciej Głowacki for stimulating conversations that changed my philosophical outlook, and to my wife Hela and brother Bartek – for everything, but first and foremost for keeping me sane through these four years.

No preface to a philosophical work dealing with epistemic and doxastic paradoxes would be complete without the following proviso: though I individually believe each sentence of this thesis to be true, I am also quite certain that a number of them are false. If this number is small, it is definitely due to all the wonderful people mentioned here; it goes without saying that the entire responsibility for this number being non-zero is mine.

Abstract

This dissertation provides an exploration of Moore's Paradox, which is the philosophical puzzle (noted first by G.E. Moore) posed by sentences of the form "*p*, but I don't believe that *p*" and "*p*, but I believe that ~*p*". Such sentences seem infelicitous to assert and irrational to believe, even though they are meaningful and consistent – therefore, their infelicity or irrationality constitutes a philosophical puzzle and invites theorizing concerning the limits of coherent speech and thought. The dissertation aims to establish theoretical ramifications for solving Moore's Paradox both as a problem concerning assertion and belief and eventually – suggest one such solution.

The dissertation is organized into five chapters and a conclusion: the first chapter introduces necessary terminology and outlines the argumentative strategy of the further chapters; the second and third chapters discuss the sources of infelicity judgments concerning Moorean speech; the fourth and fifth concern irrationality of Moorean beliefs.

Chapter 1. starts with introducing the paradox and the necessary philosophical terminology, and a critical discussion of the two historically important approaches to it: G. E. Moore's and Ludwig Wittgenstein's. Based on observing the flaws of these two accounts, I argue that a successful solution to Moore's Paradox needs to simultaneously account for its problematic nature in both speech and thought. Further on, I formulate the guiding hypothesis of the dissertation based on Sydney Shoemaker's (1995) *Priority of Belief* thesis, according to which a successful explanation of Moore-paradoxicality should be provided in belief terms, while the infelicity of Moorean assertions can be accounted for by proposing that assertion as a speech act is minimally subject to a normative expectation that it expresses a rational belief of a speaker. In the end, the chapter discusses methodological issues related to the study

of the paradox, such as the justification of intuitive infelicity and irrationality judgments, defining Moore-paradoxicality and the relationship between sentences and their propositional content.

Chapter 2 investigates contemporary theories of assertion in relation to the rational belief constraint introduced in Chapter 1. and their capacity to handle Moorean infelicity data. The chapter critiques various non-doxastic accounts of assertion from theorists such as Weiner (2005), Douven (2006), and Lackey (2007), finding them insufficient in addressing Moorean infelicity. It then evaluates alternative theories proposed by Williamson (2000), Kvanvig (2009), and Stanley (2008), which align with the hypothesis that rational belief constrains assertion. The chapter introduces and defends the discourse-sensitive account of assertoric warrant, which allows the derivation of discourse-specific norms of assertion and provides a robust explanation for the varying infelicity judgments concerning Moorean sentences with respect to the discourse in which they appear, while integrating insights from different theoretical perspectives.

Chapter 3 extends the discussion to non-assertoric speech acts, examining whether Moore-paradoxicality is confined to assertions or whether it also affects other types of speech acts. By utilizing Kent Bach and Robert Harnish's speech act taxonomy (1979) and Robert Stalnaker's "common ground" framework for analysis of pragmatic effects of speech acts (1974), the chapter argues that one can explain the infelicity of Moorean non-assertoric acts by proposing that non-assertoric speech acts could be analyzed as proposals to alter the conversational common ground. This analysis supports the broader applicability of the explanation based on the Priority of Belief hypothesis provided in Chapter 2., addressing cases involving non-assertoric speech acts such as promises, orders, apologies, or conventional speech acts.

Chapter 4 focuses on the irrationality of Moorean beliefs, comparing two strategies of explaining it: one that identifies its source in the violation of rational self-knowledge constraints (Introspectionist strategy), and another, which sees it as a consequence of

the self-defeating nature of such beliefs (Self-Defeat strategy). By formalizing the commitments of the two approaches in epistemic and doxastic logic, the chapter demonstrates that while the Self-Defeat approach has weaker theoretical commitments, it fails to account for the irrationality of all Moore-paradoxical beliefs. The analysis ultimately favors the Introspectionist approach, emphasizing the robustness of the *BB* principle defended by Introspectionists, according to which one is always in a position to rationally believe that one believes that *p* if one believes that *p*.

Chapter 5 addresses the philosophical justification for introspection principles like *BB*. The chapter defends the stronger *KB* principle, according to which if one believes that *p*, then one is in a position to know that one believes that *p*, against externalist critiques, particularly anti-luminosity arguments advanced by Timothy Williamson (2000). The chapter utilizes Alex Byrne's (2018) externalist transparency account of self-knowledge and argues that one may justifiably restrict the scope of Williamson's anti-luminosity argument by holding that introspective knowledge of one's doxastic states is not subject to margin-for-error principles. While this move allows one to uphold the *KB* principle, it also does not require abandoning externalist intuitions about knowledge and entails the rejection of the *KK* principle, according to which if one knows that *p*, then one is in a position to know that one knows that *p*.

In conclusion, this dissertation argues that Moore's Paradox can be effectively resolved by integrating a normative approach to assertion and other speech acts with robust introspective constraints on rational belief, both by demonstrating the philosophical soundness of this approach and comparing it to other solutions proposed in the literature.

List of Important Abbreviations

1. **Sentence types:**

OMP (Omissive Moore's Paradox): *"p, but I don't believe that p"*

CMP (Commissive Moore's Paradox): "*p, but I believe that ~p*"

EOMP (Epistemic Omissive Moore's Paradox): "*p*, but I don't know that *p*"

JOMP (Omissive Moore's Paradox for Justification): "*p*, but I don't have justification for believing that *p*."

COMP (Omissive Moore's Paradox for Certainty): "*p, but I'm not certain that p."*

BWK (Belief Without Knowledge): "*I believe that p, but I don't know that p"*.

OAE (Omissive Anti-Expertise): "*p if and only if I don't believe that p"*.

CAE (Commissive Anti-Expertise): "*p if and only if I believe ~p"*

IOMP (Iterated Omissive Moore's Paradox): "*p, but I don't believe that I believe that p"*

ICMP (Iterated Commissive Moore's Paradox): "*p, but I believe that I believe that ~p"*

2. **Norms of assertion and belief:**

TNA (Truth Norm of Assertion): *S* may assert *p* only if *p* is true.

RCNA (Rational Credibility Norm of Assertion): *S* may assert *p* only if *p* is rationally credible to *S*.

RTBNA (Rational to Believe Norm of Assertion): *S* may assert *p* only if it is reasonable for *S* to believe *p*.

BNA (Belief Norm of Assertion): *S* may assert *p* only if *S* believes *p*.

JBNA (Justified Belief Norm of Assertion): *S* may assert *p* only if *S* justifiably believes *p*.

KNA (Knowledge Norm of Assertion): *S* may assert *p* only if *S* knows *p*.

C-KNA (Epistemic Certainty Norm of Assertion): *S* may assert *p* only if *S* is epistemically certain of *p*.

KNB (Knowledge Norm of Belief): *S* may believe *p* only if *S* knows *p*.

3. **Principles of rational belief and knowledge:**

D: $Bp \rightarrow \sim B\sim p$; If one believes that *p*, then one does not believe that ~*p*.

BB: $Bp \rightarrow BBp$; If one believes that *p*, then one believes that one believes that *p*.

KB: $Bp \rightarrow KBp$; If one believes that p, then one knows that one believes that p.

BK: $Bp \rightarrow BKp$; If one believes that p, then one believes that one knows that p.

KK: $Kp \rightarrow KKp$; If one knows that p, then one knows that one knows that p.

4c: $BBp \rightarrow Bp$; If one believes that one believes that p, then one believes that p.

B~K: $B\sim Kp \rightarrow \sim Bp$; If one believes that one does not know that p, then one does not believe that p.

5c: $B\sim Bp \rightarrow \sim Bp$; If one believes that one does not believe that p, then one does not believe that p.

Chapter 1. How to solve Moore's Paradox: historical and methodological preliminaries

At the height of the Iran-Contra affair in March 1987, Ronald Reagan delivered a nationally broadcasted address from the Oval Office, aiming to explain the ongoing controversies and accusations concerning the actions of the President's administration regarding weapon trade with Iran's regime and funding the militant group *Contras* in Nicaragua. Although the speech is widely remembered, it is not due to the persuasive nature of Reagan's arguments. The following fragment especially became known as almost synonymous with political hypocrisy:

"A few months ago, I told the American people I did not trade arms for hostages. *My heart and my best intentions still tell me that's true, but the facts and the evidence tell me it is not*."[1]

Although at first glance the statement made by the 40th President of the United States seems inconsistent or, at best, deceitful, its philosophical analysis proves to be somewhat puzzling. To see it better, one might provide the following paraphrase of the second sentence uttered by Reagan:

(1) I believe that I did not trade arms for hostages, but I traded arms for hostages.

[1] Reagan, R. (1987, March 4). *Address to the Nation on the Iran Arms and Contra Aid Controversy* [typescript of a speech audio recording]. Ronald Reagan Presidential Library and Museum, URL: https://www.reaganlibrary.gov/archives/speech/address-nation-iran-arms-and-contra-aid-controversy-0; Accessed on: 23.09.2022.

Why is such an assertion puzzling? Though it still seems inconsistent just as if Reagan instead asserted:

(2) I did not trade arms for hostages, but I traded arms for hostages.

The contradiction in (1)'s content, unlike in (2)'s, is nowhere to be found. To see that, imagine that I believe that Ronald Reagan indeed traded arms for hostages during the 1980s, but I also believe that he believes that he did not – perhaps due to a memory loss of some kind. Therefore, I assert the following:

(3) Ronald Reagan traded arms for hostages, but he believes that he did not.

Suppose now that what I assert is actually true and Reagan, miraculously cured of his memory loss, in the next broadcast asserts:

(4) I traded arms for hostages, but (on March 4th, 1987) I believed that I did not.

It seems that both (3) and (4) are free of any absurdity or inconsistency – though their truth conditions are just the same as (1)'s. The likes of (3) and (4) seem widespread in the natural language: we often say things like 'I *was* wrong in believing that *p*', or '*She* mistakenly believes that *p*'. Therefore, it is not the content that is defective in Reagan's assertion, but the mere fact that *it is he himself who asserts the sentence* in the first-person present tense. Moreover, the prevalence of this absurdity does not seem to be limited only to *assertions*. For imagine that Reagan did not put (1) into words and kept it to himself, merely believing it. Such a scenario does not seem to make the feeling of inappropriateness go away: intuitively, *believing* a Moorean sentence is just

as absurd or irrational as asserting it. Moorean assertion and belief alike seem to disobey some deeply entrenched rules or expectations we have towards coherent speech and thought. But what rules or norms – logical, epistemic, pragmatic – get violated? What forbids us from asserting or believing something that might, after all, be true, without falling into absurdity?

This question became the topic of philosophical analysis in the late 1930s due to George Edward Moore, who posed it during his lectures, and subsequently was known (due to Wittgenstein's influential writings on the topic) as *Moore's Paradox*. Since John Williams' influential paper (1979), it has been a philosophical custom to distinguish between two forms of the 'paradoxical' sentence, known respectively as the *commissive* and *omissive* versions of Moore's Paradox (names due to Sorensen 1988):

(CMP) p, but I believe that $\sim p$.

(OMP) p, but I do not believe that p.

Although in the early studies of the paradox, the two forms had not been distinguished, it is worth noting that they remain logically independent of each other. Nevertheless, it is widely recognized that, when asserted or believed, they share a similar feeling of absurdity or inconsistency. This means that the person interested in explaining the peculiarity of (1), ought to provide an account that would also demonstrate why the following assertion or belief would be absurd:

(5) I don't believe that I traded arms for hostages, but I traded arms for hostages.

Recent years have seen a rising number of books and articles on this topic. According to Google Scholar, as of July 2024, more than 800 publications mentioning

Moore's Paradox were published only since the printing of the first monograph (Green, M. S., & Williams, J. N. (Eds.). (2007). *Moore's paradox: new essays on belief, rationality, and the first person*. Oxford: Clarendon Press.) devoted to it in 2007. Among these many provide unique takes on the problem, approaching it from the perspectives of philosophical pragmatics, epistemology, philosophy of mind, interpretation theory or even exploring its implications for metaethics (e.g., Cholbi 2009, Woods 2014), anthropology (Apter 2017) and comparative and developmental linguistics (Faller 2002, Bartha 2021). However, there seems to be no genuine agreement on which kind of solution is, or ought to count as the right one. The sheer wealth of different accounts seems to support the view, that providing *a* solution to Moore's Paradox is not as problematic as proving that it is *the* solution. This dissertation aims to fill this gap in the philosophical endeavor – to compare different existing strategies in search of the most universal and philosophically sound account.

In this chapter, I will first consider a terminological question: what is exactly 'paradoxical' in Moore's Paradox and distinguish between notions sometimes used interchangeably in the discourse concerning it: Moorean *absurdity*, *infelicity*, and *irrationality*. Further on I will provide a necessary historical context of the development of the study on and formulation of the paradox itself, and consider two of its earliest solutions: that of Moore himself and one provided by Ludwig Wittgenstein. Although both of these strategies are widely regarded as unsuccessful (as evidenced by the sheer number of articles and books aiming at providing a more satisfying solution), their failures provide an interesting insight into some of the central questions concerning the paradox. Their summary will lead me to posit the central hypothesis of this dissertation – that an adequate and unified account of Moore's Paradox needs to be spelled out in terms of the irrationality of associated beliefs. This will be introduced as a Priority of Belief Thesis (Priority), originally proposed by Sydney Shoemaker (1995; see: Williams 2013a). Priority will then be critically evaluated with respect to the existing literature, which will allow me to introduce the general plan of the dissertation

and its subsequent chapters. I will end this introductory chapter with a few methodological remarks, presenting the general method and fundamental assumptions I will use further on.

1.1. Terminology

According to the standard definition, a *paradox* is "an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises" (Sainsbury 2009, p. 1) or "an inconsistent set of propositions, each of which is very plausible" (Lycan 2010, p. 618). As such, paradoxes constitute a large and diverse family of different arguments, from seemingly counterintuitive consequences of mathematical probability theory (such as "Monty Hall problem"), through famous Zeno's arguments against the possibility of motion, plurality, and time up to semantic and logical antinomies such as the Liar or Russell's Paradox. Yet, as the very absurdity of asserted or believed Moorean sentences is neither an argument nor a set of plausible propositions, it seems that the term "paradox" is, in this case, quite loosely applied and some have preferred to use the notion of "Moore's observation" for this reason.

If we look closer, we may, however, easily locate the reason why this linguistic/doxastic peculiarity gets to be called 'paradoxical'. George Edward Moore himself characterized the paradoxicality[2] of his finding in the following manner: "It is a paradox that it should be perfectly absurd to *utter assertively words* of which the *meaning* is something which may quite well be true – is not a contradiction" (Moore

[2] Moore himself refrained from calling his observation 'paradoxical' in his early expositions, preferring to speak of the 'absurdity' of Moorean constructions. The term "Moore's *paradox*" was probably introduced by Wittgenstein in his paper presented at the meeting of the Cambridge University Moral Sciences Club in October 1945. The cited article was a response Moore provided to this paper at the club's meeting two months later, the first and the only one where he calls his invention 'a paradox'.

1993, p. 209). In this picture, "Moore's Paradox" itself seems to be that OMP and CMP stand as intuitive counterexamples to the intuitive principle, according to which one may always, without any "absurdity", believe and assert anything that is true. Though the words "absurd" and "absurdity" got deeply entrenched in the debate on Moore's Paradox, it will be more useful to characterize it as either "infelicity" in the case of Moorean assertions and "irrationality" in the case of Moorean beliefs. These are both technical terms with a more-or-less well-defined domain of application, unlike "absurdity", which is both loosely applied and emotionally loaded. Roughly speaking, both terms denote a violation of a norm we hold speech acts and thoughts to – felicity conditions of speech acts and norms of rational thought. There may be something "absurd" in following a norm or nothing "absurd" at all in disobeying it, depending on how one uses the term. Due to this vagueness, I take characterizing Moore-paradoxicality in terms of felicity of speech acts and rationality of beliefs to be a permissible terminological decision. *Moore's Paradox*, in light of Moore's suggestion, takes then the form of the following, "falsidical" (Quine 1962) paradox:

MOORE'S PARADOX:

(a) In typical circumstances[3], *p* may be rationally believed and felicitously asserted if it is true.

(b) Moorean sentences (e.g., OMP and CMP) may be true.

(c) [from (a)-(c)] In typical circumstances, if Moorean sentences are true, they may be rationally believed and felicitously asserted.

(d) In typical circumstances, Moorean sentences cannot be rationally believed and felicitously asserted, even if they are true.

[3] This provision is needed to exclude certain isolated cases in which (it seems) OMP and CMP *can* be asserted and believed. I discuss them shortly in 1.5.2.

Propositions (a)-(d) are jointly inconsistent and separately intuitively plausible, and as such constitute a paradox in the traditional sense. The premise (d) here just expresses Moore's observation. Premise (c) is justified by (a) and (b), where (b) is demonstrably true. It is, then, the intuitive principle (a) that is plausibly responsible for the rise of a contradiction. The genuine solution to Moore's Paradox then should explain the infelicity of Moorean assertions and irrationality of Moorean belief by convincingly showing us what, besides truth, constrains rational thought and speech. The two central questions posed by the paradox can be then plausibly formulated as follows:

(Q1) What condition *C* constrains felicitous assertion of some sentence *p* and is such that Moorean sentence is typically not *C*?

(Q2) What condition *C′* constrains rational belief in some sentence *p* and is such that Moorean sentence is typically not *C′*?

The purpose of this dissertation would be then to identify and defend such constraints and demonstrate how they might be applied to other issues in the philosophy of language and epistemology.

1.2. Historical background and two early accounts of Moore's Paradox

It is an uneasy task to provide an accurate description of the actual birth of Moore's observation, but, as noted by Roy Sorensen, it seems to be distinctively "a child of analytic philosophy" (Sorensen 2007, p. 38). It is indisputable that it was first noticed and discussed by Moore at the beginning of the 1930s, and as such started to

attract the attention of many philosophers visiting or studying at Cambridge in those days. Moore's first extensive discussion of the paradox may be found in notes to one of his lectures on metaphysics on the Michaelmas Term of 1932, while the first formulated Moore-paradoxical sentence can be found back in 1929 during his discussion of Bertrand Russell's account of the distinction between knowledge and true belief provided in *The Analysis of Mind*. Moore revisited this topic several times during his lectures throughout the 1930s; it seems that the problem gained attention after his remarks about it in 1936, which led to the first printed mention of Moore's Paradox in Margaret MacDonald's "Induction and Hypothesis" (1937) and later discussion focused solely on the problem in Arthur MacIver's "Some Questions About 'Know' and 'Think'" (1938). The latter of these articles became a point of reference for other philosophers and the final departure of Moore's Paradox from the Cambridge closed circles into the world, as evidenced by Langford's mention of this problem in his article devoted to Moore's account of analysis (1942).

Moore himself published only two short discussions of the paradox during his lifetime: in his reply to Charles Stevenson in the volume of *The Library of Living Philosophers* devoted to Moore's philosophy (1942) and in his discussion of Russell's theory of descriptions in an analogous volume devoted to Russell (1944). In both of these articles, the paradoxical sentence plays only a secondary role as an example introducing Moore's distinction between *asserting* and *implying* a proposition. There is evidence, however, that Moore thought of the paradox as more than a mere example. In the scripted notes for his lectures, the paradox reemerges frequently and nearly always in connection to a wider discussion on the topic of the relationship between epistemic terms such as “knowledge”, “belief” “certainty” and epistemic modals. It should come as no surprise, that the first discussion of the paradox Moore considered publishing was a part of his "Certainty" lecture delivered at the University of

California in Berkeley in 1941[4]. The paper was then again presented in the shortened version on October 26th, 1944 at the Moral Sciences Club meeting chaired by Wittgenstein, who wrote a famous letter to Moore after the lecture:

"Pointing out that "absurdity" which is in fact something *similar* to a contradiction, though it isn't one, is so important that *I hope you'll publish* your paper. (…) In a word it seems to me that you've made a *discovery*, and that you should publish it." (McGuinness 2008, p. 365)

Wittgenstein was then presumably unaware of Moore's discussions of his absurdity in his (1942, 1944); probably due to this ignorance and newly found passion, we owe the extensive discussion of the paradox in *Philosophical Investigations* and *Remarks on the Philosophy of Psychology*. Wittgenstein delivered a paper on the issue in October 1945 at the Moral Sciences Club; it was the first time when Moore's observation was dubbed a 'paradox', as evidenced by the minutes of the Club, where it is explicitly referred to as "Professor Moore's paradox: 'P, but I don't believe that P'"[5]. The last discussion of it by Moore was delivered as a reply to Wittgenstein's paper on November 29th, 1945 (it was published posthumously as [Moore 1993][6]).

---

[4] This part was not included in the final manuscript published in 1959 in the collection "Philosophical Papers". Brian McGuinness (2008, pp. 365-366) hypothesizes that some of the parts of this paper mentioning the paradox were prepared in 1944 for the presentation at the Moral Sciences Club (which may be corroborated both by his reference to distant "friends in America" as well as analogous entry in his *Commonplace Boo*k [1962, pp. 277-279], dated 1944), while other in 1941 as an earlier draft of the Berkeley lecture (which, according to McGuiness, was written on American paper). Therefore, one may easily see that the paradox seemed relevant for Moore to the content of "Certainty" both before and after the lecture was given and it remains unclear why he decided to leave its discussion out of the final version.

[5] *Minute Book of the Moral Sciences Club 1935-1952*, manuscript, University of Cambridge Library.

[6] Baldwin incorrectly dates there the manuscript to 1944 in his comments.

Probably due to Moore's discussions of the problem, the paradox was also picked up by, among others, Max Black (1952), John Austin (1940/1979[7]), Yehoshua Bar-Hillel (1946, 1954), Norman Malcolm (1950/1975, pp. 16-21), and Daniel O'Connor (1948), who all either had direct contact with Moore in Cambridge or reacted to Moore's writings. Many of the analyses in this tradition followed Moore in saying that the source of Moorean absurdity is to be found in what the speaker implies, though not explicitly asserts, and focused mostly on the psychological dimension of communication. A distinct way of approaching the problem which found its popularity in the 1950s was influenced by Wittgenstein's treatment of the puzzle present in *Philosophical Investigations*. His view, contrary to Moore's, does not treat the paradox as a mere psychological puzzle, but rather as a problem springing out from the "deep grammar" of the phrase "I believe". As such, it was also picked up by other Wittgensteinians (see, e.g., Ring, Linville 1973, 1991; Goldstein 1988, 1993).

These two approaches seem to dominate the discussion on the paradox up until the 1980s and even now are, by some, thought to be the "received view"[8]. Therefore, I will now discuss both Moore's and Wittgenstein's accounts – partly because of their historical importance and partly because their widely recognized flaws are instructive when it comes to preparing some starting desiderata that need to be met by any satisfactory account of Moore's Paradox.

---

[7] Austin's paper in question, „The Meaning of a Word", was published only posthumously. The first version of it was presented at the Moral Sciences Club in 1940 – the discussion of the paradox was either a part of it during the presentation, or perhaps was added after the discussion (Moore chaired the meeting in question).

[8] See, e.g., entry on Moore's Paradox in Michael Clark's *Paradoxes from A to Z* (2012, pp. 136-140).

1.2.1. Moore's account: asserting and implying

The first of the proposed approaches to this paradox is that of Moore himself. While the formulation of the paradox was brought by epistemological considerations such as the relationship between knowledge and true belief and the logic of the phrase "it's certain that *p*", it seems evident that Moore came at the solution he was satisfied with by the end of 1930s. Around then, in "Reply to My Critics" (1942) and "Bertrand Russell's Theory of Descriptions" (1944), he started to use his observation as a primary example of his distinction between asserting and implying[9], which he employed to deal with other philosophical and linguistic puzzles. In the first article, he writes:

"[T]o say such a thing as "I went to the pictures last Tuesday, but I don't believe that I did" is a perfectly absurd thing to say, although *what* is asserted is something which is perfectly possible logically: it is perfectly possible that you did go to the pictures and yet you do not believe that you did; the proposition that you did does not "imply" that you believe you did — that you believe you did does not *follow* from the fact that you did. And of course, also, from the fact that you say that you did, it does not follow that you believe that you did: you might be lying. But nevertheless your saying that you did, does *imply* (in another sense) that you believe you did; and this is why "I went, but I don't believe I did" is an absurd thing to say." (Moore 1942, p. 543)

In the second article, the case is presented as follows:

"[I]t is absurd to say such a thing as 'I believe he has gone out, but he has not'. This, though absurd, is not self-contradictory; for it may quite well be true. But it is absurd because, by saying 'he has not gone out' we imply that we do not believe that he has gone out, though we neither assert this, nor does it follow from anything we do assert." (Moore 1944, p. 175)

[9] The distinction was coined earlier by Moore in his *Ethics* (1907, pp. 125-127).

From the look of it, we have then two different explanations of the commissive and omissive cases – according to the first quote by saying *p* we imply that we believe that *p* (this is applied to the omissive case), according to the second we also imply then that we do *not* believe that *~p* (commissive case). Then we may look uniformly at the paradox as arising from the apparent conflict between the "implication" of the assertion of *p* and the proposition expressed by "I don't believe/believe that not *p*". This would explain the apparent absurdity in a "contradiction-like" manner without positing an actual contradiction in what the speaker asserts, therefore fulfilling the general Moorean characterization of the paradox.

The crucial question in evaluating the success of Moore's solution is, of course, what exactly he means by "implying". Though simple, the question is surely non-trivial, as one should remember that a modern interpretation of "implying" in terms of Gricean conversational implicatures would be anachronistic (though it will be discussed in the next subsection), while Moore himself does not explicitly define this notion and seems to be providing conflicting characterizations of it in different works. I think that the fairest way of interpreting Moorean sense of "imply" would be to take a look at his last paper on the paradox as well as other uses of the term that he finds analogous[10].

Moore's first printed mention of his observation in his "Reply to My Critics" mentions the paradox in his discussion of Charles Stevenson's emotivism. As he writes:

"[I]f you assert that Brutus' action was right, you imply but don't assert that you approve of Brutus' action. In the first case, that you do imply this proposition about your present attitude, although it is not implied by (i.e., does not follow from) what you assert, simply arises

[10] For a more detailed analysis of Moore's meaning of 'implying' see Atlas 2005, pp. 225-229.

from the fact, which we all learn by experience, that in the immense majority of cases a man who makes such an assertion as this does believe or know what he asserts: lying, though common enough, is vastly exceptional" (Moore 1942, p. 543)

A similar analogy and explanation are offered in "Bertrand Russell's Theory of Descriptions", where he discusses the question of why the sentence "'At least one person is a King of France' [Z] does not mean that at least one person is a King of France" seems self-contradictory:

"The absurdity I mean arises from the fact that when we use expressions to make an assertion, we imply by the mere fact of using them, that we are using them following established usage. Hence if we were to assert 'Z does not mean that at least one person is a King of France' we should imply that Z can be properly used to mean what, on the second occasion on which we are using it, we are using it to mean. And this which we imply is, of course, the contradictory of what we are asserting. We imply it, by using this language to make our assertion, though we do not assert it, nor is it implied (i.e., entailed) by what we do assert. To make our assertion by the use of this language is consequently absurd for the same reason for which it is absurd to say such a thing as 'I believe he has gone out, but he has not'." (Moore 1944, p. 175)

Therefore, according to Moore, the following sentences should be regarded as infelicitous for similar reasons:

(6) I went to the pictures on Thursday but I don't believe that I did.
(7) I believe he has gone out, but he has not.
(8) Brutus' killing of Caesar was wrong, but I don't disapprove of it.
(9) 'At least one person is a King of France' does not mean that at least one person is a King of France.

It seems then that Moore's notion of implying (or at least how he employs it in the 1940s) is very wide in scope and refers to the general psychological association by the community of language users of certain types of attitudes with certain types of

utterances. When we hear someone uttering "Murder is wrong", we expect them to disapprove murders; when we hear someone uttering sounds which remind us of a sentence of familiar language, we assume that they just used this very sentence; and, *per analogiam*, when we hear someone saying *p* in the declarative mood, we assume that they believe that *p* and don't believe *~p*. The picture of "implying" arising from this text is therefore descriptive, not normative – it is based not on norms of when one *must* or *should not* assert a sentence, but rather on inductively justified laws connecting linguistic and non-linguistic behavior.

In his critical response "Analysis of 'Correct Language'" (1946) Yehoshua Bar-Hillel provides some clarification of Moore's analysis by introducing the term "pragmatically induces" instead of Moorean loose "implies". According to Bar-Hillel's definition, the utterance A_1 pragmatically induces A_2 if and only if there exists some set of highly confirmed pragmatical laws (which Bar-Hillel defines as a law which "refers to users of language, roughly stated" [1946, p. 334]) which with conjunction with A_1 logically entails A_2. If we agree with Moore that the following two generalizations are in fact 'pragmatical laws' (see Bar-Hillel 1946, pp. 335-338; also: Grant 1958):

(*B*~) If *A* is human, and *A* utters '*p*', then *A* does not believe that *~p*.

(*B*) If *A* is human, and *A* utters '*p*', then *A* believes that *p*.

then we can explain that *A*'s utterance of the OMP or CMP sentences pragmatically induces their contrary. Similarly, according to Moore, we may explain the apparent problems of asserting (8) and (9) by postulating other pragmatical laws[11]:

[11] I will leave here the discussion of the accuracy of other pragmatical laws postulated by Moore. The reader might be referred to the first part of Bar-Hillel's paper for an extensive (and in my opinion – accurate) critique of Moore's account related to (L); an interesting discussion of cases similar to (8) and the pragmatic law (M) might be found in Woods 2014, where it is also argued that assertions alike to (8)

(M) If *A* is human, and *A* utters '*φ*-ing is wrong', then *A* disapproves of *φ*-ing.

(L) If *A* is human, and *A* utters '*p*', and *p* is a sentence in language *L*, then *A* said that *p* in *L*.

Bar-Hillel's clarification is useful since it allows us to highlight two crucial features of Moore's solution. First of all, it relies on empirical assumptions connecting speakers' linguistic behavior with their attitudes. It is important to understand that by "linguistic behavior" we mean just uninterpreted sounds, inscriptions, or gestures produced by a linguistic agent, for there is nothing in their observable behavior that indicates that they are a*sserting* something (despite, perhaps, a tone of voice or air of confidence). Secondly, and even more importantly, it seems initially incorrect to say that *the speaker* implied anything – pragmatic implication in the above sense is *not* an act of the speaker, but rather a property of the utterance itself. Only if, after Black (1952), we assume that such pragmatical laws of the cited form constitute certain widely accepted conventions present in communication, we might say that the speaker *implies* something by intentionally exploiting these laws. "[A] man who says 'p' while disbelieving *p* is breaking a rule about a conventional sign", says Black, "like an unmarried woman who wears a wedding ring" (Black 1952, p. 32). Although someone might just unknowingly wear a wedding ring as a piece of expensive and stylish jewelry and not "imply" knowingly anything by their appearance, when they are aware of associated conventions, they might be rightly said to exploit them to achieve their goals and properly "imply" that they are married.

With these two features in mind, we might proceed to the evaluation of Moore's strategy. Firstly, the strategy would be a non-starter if the pragmatic laws mentioned

---

importantly differ from traditional Moore-paradoxical assertions, while the opposite conclusion is argued in Cholbi 2009.

above were not well confirmed and believed so by the linguistic community. For if the violation of these laws is sufficiently common, then there would be no point in expecting that the speaker's utterance would imply anything by (B) or (B~), and hence, there would be nothing "absurd" in their behavior. The point that these laws are not sufficiently well confirmed is one of the main and common objections against Moore's solution, which, in turn, had been colorfully described as "a naïve underestimate of the frequency of lying" (Sorensen 2007, p. 48). The main problem of the Moorean strategy, however, lies somewhere else. We might try to take (B~) and (B) for granted[12] and see, whether the Moorean strategy works even if we assume that its premises are true.

According to a second crucial feature of Moorean explanation, the "implication" cannot, in a relevant sense, be said to be an action of the speaker, but rather a property of the utterance. As mentioned above, the speaker may be said to imply *p* by saying *q* only if an utterance of *q* pragmatically induces *p* and if a speaker knows that the associated pragmatical laws exist and wishes to exploit them. This should lead us to think that in the absence of some deceitful intent, the speaker uttering a Moore-paradoxical sentence is not guilty of any violation, but merely unaware of existing conventions and pragmatical laws. Let us then consider the following dialogue:

(Dialogue 1) A: Brutus' killing of Caesar was wrong.

B: So, you believe it?

A: I did not say that, in fact, I don't believe it. It does not change the fact that it was wrong, however.

[12] Note that, *contra* Sorensen and other critics, at least (B) seems to have genuine support in empirical communication studies; see, e.g., Serota, Levine 2015.

Let's assume that A is sincere and has no intent to deceive B. Is A then simply unaware of related conventions or is his utterance an exception to (B)? Or can A be criticized as incoherent? It seems that even with all our assumptions, two last of A's assertions form a Moore-paradoxical pair, which infelicity remains unexplained. Similarly, the infelicity of such an assertion does not go away if we know that our interlocutor is lying (and hence (B) and (B~) do not apply to them).

The pragmatical laws as described by Bar-Hillel and Moore have also several different other consequences, which may further elucidate why such an account of Moore's Paradox is insufficient. Consider the following utterance:

(10) A (*angrily shouting*): I am calm, goddamn it!

As Bar-Hillel (1946, pp. 334-335) notices, such an utterance can also be described as "absurd" in Moore's sense (since we might formulate a well-confirmed pragmatic law connecting the tone of voice and the use of expressive language with the excitement of the speaker[13]). But is it *comparably absurd* to Moore-paradoxical assertions? Certainly not – for it does not indicate that the speaker is incoherent, just, perhaps, that they are not fully controlling themselves. The fact that Moore's account *de facto* equates the two is strong evidence that it fails to explain the peculiar infelicity of Moorean assertions.

[13] Note also that in this example the resolution of the apparent absurdity is fairly easy – in most cases we would say that A is excited, following the prescription of the pragmatical law rather than the semantic content of their exclamation (unless one was joking, acting, or intentionally misleading). It is however not clear how we should approach a person uttering a Moore-paradoxical sentence: do they believe that *p* (as implied by their utterance of *p*) or do not (*per* the semantic content of the second conjunct)?

The problems we face while providing an account based on inductively established laws of human linguistic behavior seem to have their source in the fact that pragmatical laws as envisaged by Moore and Bar-Hillel necessarily associate psychological consequences with utterances rather than speech acts. Therefore, it is both overly general, requiring complicated adjustments to account e.g., for the difference between the apparent infelicity of genuine assertion of OMP or CMP and the felicity of uttering OMP or CMP as part of a joke, an act, and so on, as well as not philosophically satisfying. Moore's original analyses seem to answer rather a different question than we were originally presented with: the only thing they can account for is the question of why *utterances* of OMP or CMP *sound bad*, but not the more general and fundamental question of why *assertions* of OMP and CMP *cannot be felicitously performed*.

The general lesson from the inadequacy of Moore's account can then be summed up as follows: to resolve Moore's Paradox in speech, one would be ill-advised to restrict themselves to merely empirical considerations concerning linguistic behavior, omitting the normative expectations language users have towards speakers. It is plausible that to do so, we need theoretical reflection on what norms[14] govern the practice of asserting and performing other speech acts.

[14] Many were tempted to treat Moore and Black as precursors of normative treatments of assertion, especially in the "knowledge norm" tradition (e.g., Williamson 2000, p. 252, f. 6 and his followers), citing Moore's one-off stronger claim from the *Commonplace Book* that "by asserting *p* positively you *imply,* though you don't assert, that you know that *p*" (1962, p. 277). I take this section to demonstrate that this association is definitely false.

1.2.2. Neo-Gricean interpretation of Moore's strategy

Before moving on to Wittgenstein's take on the problem, I shall return quickly to the temptation of straightforwardly understanding Moore's sense of "implying" along the lines of Gricean theory of implicatures. This theory may perhaps better accommodate the normative demand if we think of Grice's maxims as norms regulating the conversational practice, and be the last resort of a "purely pragmatic" strategy of explaining its puzzling character. Although it is almost definitely not an adequate way of *interpreting* Moore, this kind of solution to the paradox gained some popularity in the 1970s and 1980s (Martinich 1980, Levinson 1983, Tokarz 1993) and (as we shall see in Chapter 2.) remains popular today; it is then important to discuss it.

The "Neo-Gricean"[15] interpretation of Moore's remarks might look as follows. Roughly, one may say that when Moore speaks of "implying" one can understand him as meaning "conversationally implying" in Grice's sense, that is – following from conversational maxims accepted by conversation participants acting under the cooperative principle. According to the Maxim of Quality, a speaker should "try to make their contribution one that is true", and therefore obey the following sub-rules (Grice 1975/1989, p. 27):

(*Quality-1*) Do not say what you believe to be false.

(*Quality-2*) Do not say that for which you lack adequate evidence.

If a speaker asserts p, we then take him to conversationally imply that they do not believe $\sim p$ (by the first rule) and that they possess adequate evidence for the truth of p (by the second rule) – so, presumably, they may be interpreted as believing p. This

[15] As I note at the end of this section, this explanation was explicitly denounced by Grice himself – therefore "Gricean" comes with a "neo-" prefix.

allows us to say that asserting the first conjunct of CMP and OMP[16] we conversationally imply something that is directly contradictory to the semantic content of the second conjunct. On such account, Moore's Paradox in general is a blunt contradiction between *what is said* and *what is implied* by the speaker.

This solution, however, may be easily refuted by observing that one of the defining features of implicatures distinguishing them from, for example, presuppositions is they can be canceled[17] (Grice 1981) and non-redundantly made explicit (Sadock 1978). That is, if I assert:

(11) Some students have passed the test.

I am also allowed to cancel the scalar implicature of (11) derived from the Maxim of Quantity, that is to felicitously say:

(11a) Some students have passed the test, *but I don't mean to imply that not all students passed the test.*

or to strengthen my assertion as to encompass the implied content without feeling of redundancy:

(11b) Some, *but not all*, students have passed the test.

---

[16] Martinich appeals to a more general statement he deduces from the Maxim of Quality: "A speaker who asserts that *p* conversationally implies that he believes that *p*" (1980, p. 224). It is, however, unclear how this should account for the absurdity of CMP, since the fact that the speaker believes that *~p* is not *logically* contradictory with the fact that he believes that *p*. It is, therefore, more promising to stick with the original Gricean formulation (which also interestingly coincides with the two pragmatical laws featured in Moore's explanation).

[17] This claim has been disputed by e.g., Carston (2002), Bach (2006), and Lauer (2014); for discussion and arguments for the contrary see Zakkou (2018), and Puczyłowski (2020).

This does not seem to be true, however, of pairs of assertions with associated first-person belief reports. Note that canceling supposed Quality implicature simply *results* in Moorean infelicity[18]:

(11c) Some students have passed the test, but I don't mean to imply that I believe that; in fact, I don't.

while (unlike in 11b) strengthening one's assertion with making the content of such implicature seems redundant:

(11d) Some students have passed the test and I believe that.

Unless we abandon the standard tests for implicatures, the explanation of Moore's Paradox in these terms is hopeless. Interestingly, Grice himself rejected the solution in question and in general the thought that the speaker *conversationally implies* that they possess the relevant belief or disbelieve its negation:

"On my account, it will not be true that when I say that *p*, I conversationally implicate that I believe that *p*; for to suppose that I believe that *p* (…) is just to suppose that I am observing the first maxim of Quality on this occasion. I think that the consequence is intuitively acceptable; it is not a natural use of language to describe one who has said that *p* as having, for example, "implied", "indicated", or "suggested" that he believes that *p*; the natural thing to say is that he has expressed (…) the belief that *p*." (Grice 1978/1989, p. 42)

This may indicate that while there is a lot to be drawn from speakers' words about their states of mind *via* the assumed set conversational maxims, Moorean sentences in a sense lie too close to the negation of the Maxim of Quality itself. Later in Grice's paper follows an analysis of indicative mood and the speech act of assertion, which support the hypothesis that Grice thought assertions do not imply the speaker's

[18] This point is also made by Baldwin (1990, p. 228) and Levinson (1983, p. 105, f. 7).

belief in their content but rather directly *express* that belief[19]. The solution that tries to work with a similar intuition, as proposed by Ludwig Wittgenstein and his philosophical descendants, will be the topic of the next section.

1.2.3. Wittgenstein's account: belief and expression

As noted in the opening paragraphs of this section, the second philosopher to develop their account of Moore's Paradox was Ludwig Wittgenstein. The general outline of his thoughts on the Moore-paradoxical assertions and the way they differ from Moore's own can be found even in the letter to Moore quoted above, written just a few days after his first encounter with the paradox:

"To call this, as I think you did, "an absurdity for *psychological* reasons" seems to me to be wrong, or *highly* misleading. (…) By the way, don't be shocked at my saying it's something "similar" to a contradiction. This means roughly: it plays a similar role in logic. You have said something about the *logic* of assertion. Viz: It makes sense to say "Let's suppose: p is the case and I don't believe that p is the case", whereas it makes no sense to assert "p is the case and I don't believe that p is the case". This *assertion* has to be ruled out and *is* ruled out by "common sense", just as a contradiction is. And this just shows that logic isn't as simple as logicians think it is." (after McGuinness 2008, p. 365)

As witnessed here, Wittgenstein thought that we should not approach Moore-paradoxical assertions from the point of empirically informed pragmatics and

[19] Another way to read this remark is that Grice thinks of Quality as distinct *normative* constraint on assertion in a way similar to the one which I present in the next chapter (Benton 2016 interprets Grice this way, noting that only Quality refers explicitly to assertoric speech). Given the positioning of Quality among other maxims and its supposed derivation from the Cooperative Principle, I think that this interpretation is at least non-standard (though welcome if true).

psychology, but rather rules governing assertion and, especially, first-person belief avowals.

The need to provide a satisfying account of Moorean absurdities lasted long in his mind. One may find extensive passages devoted to it in *Remarks on the Philosophy of Psychology* and *Last Writings on the Philosophy of Psychology,* which served as a preparatory work for the second part of *Philosophical Investigations* (written between 1946 and 1949) in which a short, three-page section *X* is devoted solely to an explanation of the paradox[20]. The main intuition behind Wittgenstein's approach to the absurdity of OMP and CMP as expressed there seems to be that usually assertions of the form "I believe that *p*" serve as, perhaps a bit hesitant, assertions of *p*. In the letter to Moore, Wittgenstein similarly remarks:

"If I ask someone "Is there a fire in the next room?" and he answers "I believe there is" I can't say: "Don't be irrelevant. I asked you about the fire, not about your state of mind"" (after McGuiness 2008, p. 365).

In line with this observation, it is *prima facie* plausible that the puzzling Moorean assertion of the CMP form is *actually* a form of a contradictory assertion "*p*, but *~p*". This is reinforced by Wittgenstein's elaboration when he stipulates the possibility of the language in which "mistakenly/falsely believes" is a single verb. According to the Austrian, such a verb "would not have any significant first-person present indicative" (Wittgenstein 1999a, p. 191); this points out the fact that although we may try to *use* the verb "believe" to ascribe a mistake to our present selves, such use is simply "ungrammatical" (in Wittgenstein's sense of "deep grammar"). Similarly, Norman

[20] As reported by Monk, he was "particularly satisfied" with the final result (Monk 1991, p. 560); hence, I focus on Wittgenstein's approach as presented in *Investigations*, ignoring some more subtle considerations regarding the variants of his main idea as featured in *Remarks* and *Last Writings.*

Malcolm in his Wittgensteinian interpretation reports his intuition that such sentences seem to him plainly unintelligible (1995, p. 195).

Wittgenstein's solution can be thought of as a form of expressivism about belief reports. A report "I believe that *p*" is an assertion of *p* (a "hesitant assertion", Wittgenstein acknowledges, but not an "assertion of hesitancy"; see Wittgenstein 1999, p. 192) in virtue of *expressing* our belief that *p*, which is a property it shares with a plain assertion of *p*. As he writes in "Remarks…" (1998, §472): "With the assertion *It's going to rain* one expresses belief in that just as one expresses the wish to have wine with the words *Wine over here!*". One may think here of an analogy with another well-known Wittgensteinian analysis: by saying "I know that I am in pain", the agent is not stating anything more than by saying "I feel pain", or even by simply exclaiming "Ouch!" or cursing (1999, §246). Verbs such as "being in pain" (and, by analogy, "believing" or "wishing") serve primarily to ascribe mental states to others, while their use in self-ascriptions is superfluous and merely expressive.

As to any form of expressivism, the following objection easily comes to mind, known in a general form as a Frege-Geach Problem (Geach 1965[21]). Suppose I assert "I believe that *p*". It seems that I may then rightfully apply logical operations to it and deduce, e.g., that "Someone believes that *p*" by existential generalization. But if Wittgenstein is right in that I am not asserting anything about myself by my belief self-report, then such a deduction is unjustified, for the truth of *p* does not entail that someone believes that *p*. If I am, however, justified in making such inference, it seems

[21] In Geach's original presentation (1965, pp. 462-463), he discusses a view ascribed to Austin, according to which first-person knowledge ascriptions in the form of "I know that *p*" are simply guarantees, expressing high confidence in *p*'s truth, which corresponds well with Wittgenstein's proposal discussed here.

that there is something more to such an assertion than a more elaborate or cautious way of asserting *p*.

It is uncertain how Wittgenstein himself would respond to such a line of reasoning. There are some reasons to think that he would not be bothered by such a result and maintain that there is no justification for this inference to be valid. As he writes in Section *X*, assuming that the verb "believe" in the past-tense assertion "I believed that *p*" (or in the supposition "Suppose that I believe that *p*") has the same meaning as the present-tense "I believe that *p*" may be simply a false reasoning by analogy, as saying that "$\sqrt{-1}$" must mean to –1 the same as "$\sqrt{1}$" does mean to 1 (1999, p. 190). But this response is hardly convincing: although it might explain why our intuitions seem to drive us towards a conclusion that both assertions differ only with respect to tense, but not their meaning, it does not provide any useful and principled way of explaining why Frege-Geach inference is invalid (as providing a counterexample would, for example, and which might be easily done in case of the square root of –1 and 1). A similar objection could be applied to a more modest Wittgensteinian response, pointing out that there are some cases in which "I believe that ~*p*" could be used as a hypothesis based on the observation of one's behavior (perhaps by saying "Judging from what I say, *this* is what I believe." [1999, p. 192]), and could play a role in such inferences. The Frege-Geach inference is not limited to such cases and seems to work as well with assertions that are not based on such external justification.

The problem, as of yet, seems to be that merely noticing that the verb "believe" is used differently in the first-person present tense than in normal folk-psychological ascriptions used to explain and predict the behavior of others cannot constitute a radical divorce in meaning between these two patterns of use. An amendment of this expressivist strategy, proposed e.g. by Jane Heal (1994) and Dorit Bar-On (2004), may try to account for these shortcomings of Wittgenstein's approach by differentiating the

expressive and semantic dimensions of Moorean assertion[22]. While, as shown above, there is little to no prospect of a *semantic* reduction of belief self-ascriptions to endorsements of belief's content, one may still argue that such reduction is of an expressive, not semantic nature. Let us go back to the starting intuition that asserting "I believe that *p*" is usually taken to be a (hesitant, or "hedged" in contemporary terms) assertion of *p*. To account for it, an expressivist may try to posit that although "I believe that *p*" does not *mean* the same as "⊢*p*" (as Wittgenstein [1998, §478] would have it), it *expresses* the same mental state, that is the belief that *p*. Then, although the semantic contents of "*p* and I believe that ~*p*" and "*p* and ~*p*" assertions are different, they both express the same pair of contradictory beliefs and *this* fact may explain the peculiarity of Moorean assertions. As Dorit Bar-On claims in her neo-expressivist account of avowals, Moorean statements seem to exhibit an "expressive conflict", for the belief that *p* is my reason to both assert that *p* and that I believe that *p* (Bar-On 2004, p. 218).

A careful reader of this passage would without a doubt notice something peculiar by now. From the beginning of an explanation of Wittgenstein's position and proposed amendments I carefully worded every example of a Moore-paradoxical assertion in its commissive form, leaving omissive cases aside. It seems that such assertions may cause significant trouble: while "I believe that *p*" may be regarded as an expression of one's belief that *p*, "I don't believe that *p*" may not, at least at face value, express *a lack of belief*, because the lack of a mental state is not itself a mental state one may easily express.

[22] Though Wittgenstein himself considered and rejected a similar approach; see Malcolm 1995, pp. 199-203.

While Heal in her account simply omits the question of handling the omissive Moorean assertions, Bar-On acknowledges this worry and replies by interpreting such "lack-of-belief" avowals in the following manner:

"The more natural interpretation, I would argue, is one on which the avowal is a cautious or less blunt, way of expressing the avower's belief that not-*p*. In that reading, we get the same type of conflict as with "I believe that *p*, but not-*p*" (except that "*p*" and "not-*p*" are reversed). The more literal interpretation is one according to which the second conjunct simply expresses the avower's uncertainty or complete agnosticism regarding *p*. (A more natural way of expressing this attitude would be to avow "I'm not sure whether *p*" or "I have no opinion about *p*".)" (Bar-On 2004, p. 219)

In other words, it seems that according to Bar-On by making even such a negative self-ascription, we nevertheless express some relevant mental state, perhaps that of some less-than-certain *credence* (for the *lack of opinion* or *lack of certainty* are not, again, mental states, if lack-of-belief is not) or plain disbelief[23]. I think there is some truth to this observation, as we certainly sometimes resort to saying "I don't believe that *p*" when we disbelieve *p*. Is it however true that we *always* express an associated first-order mental state with such avowals? Let us look at the following examples:

(12) I neither believe that ~*p* nor that *p*.

(13) I neither am unsure whether *p* nor believe that *p*.

(14) I have no (positive, negative or neutral) confidence in *p* being true or false.

I take (12)-(14) to have some natural use conditions, and not be universally infelicitous; but if we agree with Bar-On's analysis, this should be impossible, and (12)-(14) just as infelicitous as CMP and OMP. For if we grant, along the lines of the first

[23] This latter option is fully endorsed by Goldstein 1993.

hypothesis, that belief disavowal expresses disbelief, then (12) expresses contradictory beliefs in ~*p* and ~~*p*, and if we allow such statements to express credence, (13) and (14) should be thought to stand in expressive conflict themselves since they supposedly nevertheless express some sort of (negative or positive) credence in *p*'s truth. Though one could in principle defend such a stance, it seems to me rather hopeless and unpromising. To give some plausibility to my claims of the felicity of (12)-(14), consider the following example: according to the knowledge I gained from skimming a Wikipedia article about it, there is no established evidence for the existence of water (nor specific evidence for the contrary) in the Sunflower (Messier 63) Galaxy. Not being an astrophysicist, I have no way of telling whether finding water there is even plausible; I would hesitate to ascribe *any* subjective probability to the statement "There is water in the Sunflower Galaxy", for I do not have the appropriate skills and expertise to do it, and going *50/50* strikes me as an irrational response, given the fact that I have no idea whether the chances of finding water there are such. If pressed, I would by all means say that I do not believe that there is water in the Sunflower Galaxy, nor do I believe that there is none; I would also assert that I have *no* degree, neither positive nor negative, of certainty (no credence at all) concerning this fact.

What do I express then by asserting: "I don't believe that there is water in the Sunflower Galaxy"? I think that the most straightforward answer – that I only express *my higher-order belief* that I do not believe that there is water in the Sunflower Galaxy, not some in-between credence – is *prima facie* the correct one. If that is the case, then if I'd assert:

(15) There is water in the Sunflower Galaxy, but I don't believe that.

I would only express that I believe that there is water in the Sunflower Galaxy and that I believe that I don't believe it. There is nothing contradictory in this fact alone

unless we posit some further principles concerning the consistency of beliefs. Without it, the expressive account fails to deliver the expected outcome[24].

Let us also notice that the expressive-conflict account of Moorean assertions as well as an orthodox Wittgensteinian picture may work only if such an assertion is a conjunction – that is, if we may derive two separate assertions expressing conflicting attitudes. What happens if we embed the Moorean content, so it is not represented by a conjunction? Consider the following cases presented by Sorensen (1988):

(16) My atheism angers God.

(17) Christmas is closer than I believe.

or the following embedding of a Moore-paradoxical sentence into a third-person knowledge report:

(18) He knows that: *p* and that I don't believe that *p*.

Although these examples may lose a bit of their striking absurdity when compared to the original OMP and CMP, they are still problematic to assert from a first-person present perspective. But the only thing that such assertions may be said to express is a singular belief of the speaker. By the very definition of "conflict", their problematic nature cannot be accounted for by expressing *two* conflicting mental states unless we will assume some background principles regarding the relation between one mental state of believing the negation of a material conditional and two attitudes

[24] One could try to resist and say that apart from my higher-order belief, I simultaneously express my *agnosticism* or a *suspension of judgment* considered as a primitive mental state concerning the existence of water in the Sunflower Galaxy. But even conceding that, we still need some background principles governing the interaction of agnostic and belief states; simultaneously being agnostic towards and believing that *p* does not constitute a contradiction in itself if 'suspension' is a *sui generis* state, just as believing that *p* and believing that one does not believe that *p* does not.

toward its antecedent and consequent. Since this remains out of the scope of Wittgenstein's or Bar-On's explanation, their theories are too weak to provide us with something more than merely an account of the linguistic inappropriateness of only some Moorean constructions.

1.2.4 Towards a unified account of Moorean irrationality and infelicity

As I have shown above, both Moore's own and Wittgenstein's accounts of the paradox have important shortcomings. What can we learn from them?

The crucial pitfall of both of these strategies results from, I take it, solely focusing on the linguistic dimension of the problem. As noted in the opening section, Moore's Paradox is a problem emerging both at the level of speech *and* thought; *believing* Moorean sentences seems irrational or peculiar just as *asserting* them elicits judgment of infelicity. The problems of both Moore's and Wittgenstein's accounts can, on a broader picture, be offered a similar diagnosis of their shortcomings: by seeking an explanation on the linguistic level, they try to make Moorean assertions similar to overt contradictions by postulating various linguistic or psychological conventions, without considering the hypothesis that what needs to be explained happens at the level of belief. A good example of this lesson may be drawn from the concluding remarks of the last section. While many, especially early accounts of Moore's Paradox tried to account for the absurdity of consecutive assertion of both conjuncts of OMP or CMP, it does not seem to capture well the unassertability of Moorean statements in other syntactic forms. One may think here also of other assertions which entail Moorean statements, such as Wittgensteinian in spirit:

(19) I falsely believe that *p*.

(20) I don't hold a nevertheless true belief that *p*.

which cannot be easily thought to express two opposing mental states at once. What is rather plausibly happening is that by (16)-(20) and, consequently, by OMP and CMP one is expressing a single belief (namely the belief that: *p and one does not believe/believes that not p*) – and what we need to do is to account for its irrationality.

While, in principle, it is perfectly possible to provide a separate account of Moorean infelicity in speech and Moorean irrationality in thought, it is only natural to seek a unified strategy of explanation. If we are to seek such a strategy, we may consider two routes of going about it. The first route is top-down: we try to figure out the roots of Moorean infelicity and then go on to explain why having a belief which content is adequately given by a pragmatically infelicitous sentence is irrational. As the examples of Moore and Wittgenstein demonstrate, providing a solution to Moore's Paradox purely at the level of speech seems problematic, or at least non-trivial. In what follows, I will take the second, bottom-up route: explain Moorean infelicity in speech in terms of the irrationality of Moorean beliefs. In the next section, I will try to clearly state the hypothesis that providing an account of the irrationality of Moorean beliefs is explanatorily prior to accounting for the infelicity of Moorean assertions.

1.3. The Priority of Belief thesis and the bridge between assertion and belief

How one should then approach the task of explaining the paradoxicality of Moorean assertions and beliefs? In the previous section, I made it clear that a purely pragmatic strategy (that is – a strategy that appeals only to rules of interpretation of speech) would likely not suffice to prove anything more than a feeling of inappropriateness related to certain linguistic constructions, which both do not exhaust the variety of speech acts which may be called Moorean as well as would not account for the paradoxicality of Moorean assertions. I also suggested that following a route from belief to assertion, not *vice versa*, may yield better and more universal

results. In a similar vein, Sydney Shoemaker argues for focusing on the irrationality of Moorean beliefs as primitive:

"What really needs to be explained is why someone cannot *believe* that it is raining and that she doesn't believe that it is, despite the fact that the conjuncts of this belief can both be true. If we can show that such beliefs are impossible, or at least logically defective, and if we come up with an explanation of this, then an explanation of why one cannot *assert* a Moore-paradoxical sentence will come along for free, via the principle that what one can believe constrains what one can assert." (Shoemaker 1995, p. 231)

The thesis put forward in this quote has been dubbed the "Priority of Belief" thesis in the literature and may be represented in the following form:

(**Priority of Belief**) The infelicity of Moore-paradoxical assertions is explained by the irrationality of Moore-paradoxical beliefs.

To give some initial plausibility to Priority, Shoemaker invokes the following "bridging" principle:

(*Bridging Principle*) One can assert only what one can believe.

If the Bridging Principle is true, it seems that Priority follows almost immediately. Of course, the significant problem here is how one should read both occurrences of "can" in Shoemaker's formulation. Are we speaking here of metaphysical or conceptual possibility, or some form of permissibility? Descriptively speaking, one *can* have beliefs even more outlandish than Moorean ones (think, e.g., of a patient with a Cotard delusion believing that they are dead); clearly, we need not claim that Moorean belief is impossible, but just irrational, and read the first "can"

normatively[25]. What about the second "can"? Is rational belief a constraint on all assertions? Of course, one need not believe what they say for their utterance to *descriptively qualify* as an assertion, for lies are also assertions, although they still plausibly conventionally express beliefs. Should we then, as with beliefs, speak here of "rational assertion"?

The case gets tricky fast, as "rational assertion" may be interpreted in a variety of ways. One way to understand it would be to take assertion to be guided by *practical* rationality principles – as a speech *act*, assertion needs to accord with such norms. But one may be practically justified in asserting something one does not believe or believes on poor epistemic grounds – for it may be practically rational for one to communicate it. For example, John N. Williams (2013a; 2023) presents a variety of different cases in which one is supposedly practically rational in expressing their epistemically irrational Moorean belief[26]. Though Williams (seconded by Littlejohn [2020]) takes similar examples to speak against Priority and the likes of the Bridging Principle in general, they are valid only insofar as we take assertion to be guided only by practical norms. For intuitively, there is still something wrong with the patient's assertion – it signifies their irrationality and unreliability, which we usually do require from other assertors.

---

[25] Though in (1988), Shoemaker seems to argue for the "impossibility" of Moorean belief, he later formulates the principle more cautiously in an endnote (1995): "[T]he principle might be better put: What can be *coherently* believed constrains what one can be *coherently* asserted" (Shoemaker 1995, p. 227, f. 1; my emphasis).

[26] Some of these cases resemble the "atypical" ones introduced by Pruss (2012), Borgoni (2015) or Fileva and Brakel (2019), in which the speaker gains evidence that they believe that *p* through third-person testimony or external observation of one's behavior; I discuss these cases briefly in section 6.2., as well as at the end of this chapter.

A better way to understand "rational assertion" would be to think not of general rules of practically rational action, but rather norms specific to assertion and the practice that constitutes it. Just as any act, an assertion may be practically rational given the agent's desires, beliefs, and intentions, but not conform or even radically break norms associated with the practice – think of a chess player, who may be practically rational in deliberately making an illegal move. The fact that one may be practically justified in asserting a Moorean sentence does not contradict the fact that such assertion still plausibly remains *improper* with respect to normative constraints specific to asserting – we may criticize the speaker for falling short of relevant standards of assertion even if they are practically rational in making it and are not insincere. To see this, consider Williamson's example of shouting "This is your train!" to someone, despite not knowing (or even fully believing) that this is, in fact, the right train (Williamson 2000, p. 256). In such a case, the urgency of the situation requires me to act fast and without proper deliberation – asserting that this is your train is a *practically rational* thing to do. But it does not demonstrate that I am thereby warranted in my assertion nor that such an assertion is proper, but only the fact that such norm may be "overridden" by other norms of practical action. I may still be plausibly criticized for asserting that this is your train if, after all, it so happens that you get on the wrong one as a result of my unwarranted assertion.

What then, if not practical rationality, warrants assertion and allows us to call it *proper*? Usual guesses among the contemporary theoreticians of norms of assertion put distinctively epistemic constraints on its content: assertion needs to be true, reasonable to believe, justifiably believed, or known by the assertor. Crucially, all of these norms require that the assertor possesses a specific, epistemic warrant for assertion. Given this fact and the appropriate connection between distinctively epistemic reasons one needs to have to hold a *rational belief* and to make a *warranted assertion*, Shoemaker's principle may be formulated in the following way:

(**Bridging Principle**) One can warrantably assert only what one can rationally believe.

The Priority of Belief thesis shall be given then the following interpretation: the infelicity of Moorean assertion is explained by the fact that Moorean sentences can never be warrantably asserted, which in turn (*via* the amended Bridging Principle) is because they cannot be rationally believed. A good explanation of Moore's Paradox in belief taken together with a sufficiently strong normative account of assertion may then be enough to immediately give us an account of why Moorean assertions are infelicitous and, more importantly, give us a unified account of what goes wrong both in Moorean thought and speech. In the context of our guiding questions (Q1) and (Q2), as introduced in section 1.1., the hypothesis can by interpreted as follows: the condition *C*, constraining warranted assertion, is that of *rational belief*, while the fact that believing Moorean sentences violates some other condition *C'* of epistemic rationality is taken to be central and explanatory for both Moorean phenomena.

In the following dissertation, I will evaluate the prospects of defending Priority of Belief, which I shall refer to simply as the Priority thesis from now on, and provide an explanation of the irrationality of Moorean belief which, via Priority, shall also yield an explanation of infelicity Moorean assertion. In what follows, I want to quickly outline the plan of attack, starting from the questions concerning the connection between rational belief and different accounts of assertion and other speech acts (part I) and finishing with the questions concerning the source of the irrationality of Moorean beliefs (part II). In the ending section of this introductory chapter, I will also offer methodological considerations and an attempt at a definition of Moore-paradoxicality which will help to distinguish true Moorean phenomena from otherwise puzzling speech acts and mental states bearing only superficial similarity to the original paradox.

1.4. Plan of the dissertation

How should we approach proving Priority and explain the irrationality of Moorean beliefs? In this section, I will outline some possible obstacles for Priority and sketch the general approach I will take to demonstrate why they are not successful. Furthermore, I will address the question of proper explication of the *C'* belief rationality constraint – that is *why exactly* Moorean beliefs are irrational and their irrationality strikes us so boldly.

1.4.1. Two questions concerning Moorean speech

The most fundamental task that I will undertake in the first two chapters concerns the scope of Moore-paradoxicality in speech – that is, what cases exactly should an account of Moorean infelicity cover and how one may explain it in a manner coherent with Priority? Why should we even presume that Moore-paradoxicality is exhibited only by assertions and beliefs, not other speech acts and propositional attitudes or mental states?

The first question concerning the explanation of Moorean infelicity concerns a choice of a theory of assertion that may supplement Priority. Aiming at even listing all considered views on how to define, characterize, or constrain the notion of assertion is probably an impossible task. However, one may try to find at least a range of views out of those proposed in the literature that may allow for a reductivist explanation of the infelicity of Moorean assertions in terms of the irrationality of Moorean beliefs that align with the Priority thesis. As noted above, I shall limit myself to those theories that present assertion as a norm-guided practice (as characterized in Pagin, Marsili 2021). Descriptive accounts that aim at characterizing this practice without referring to their normative aspect would be considered only as tools for primitive assessment of what counts as an assertion (and differentiate it, for example, from a conjecture or a

promise), while the source of infelicity judgments would be located in the speaker violating the norm pertaining to and regulating this very speech act.

In this chapter, I will argue that only sufficiently strong normative accounts of assertion can fulfill the theoretically desired role compatible with Priority, while at the same time explaining the infelicity of certain "stronger" kins of OMP and CMP:

(EOMP) *p*, but I don't know that *p*.

(JOMP) *p*, but I don't have justification for believing that *p*.

(COMP) *p*, but I'm not certain that *p*.

which are also classified by many as Moorean, though beliefs corresponding to them are not so easily labeled as irrational (see Hintikka 1962, McGlynn 2013): believing something without taking this belief to constitute *knowledge* proper, or even more so without full certainty seems ubiquitous[27]. Hence, I will postulate that the standard we hold assertions to is usually higher, and propose a way of relativizing this standard for assertion, which helps to explain discursive differences in felicity judgments regarding EOMP and COMP.

While the first question focuses on norms governing assertion and establishing a Priority-based continuous strategy of explaining the infelicity of Moorean assertions, chapter 3. will discuss how it could be extended to other, not necessarily constative, speech acts. For not only assertions had been called "Moorean", and many authors found the likes of the following constructions strikingly similar to OMP and CMP:

(21) I promise to marry you, but I will not marry you.

[27] I shall return to this question briefly in Chapter 2. and more extensively in Chapter 5.

(22) Please, close the window, but I don't want you to do it.

(23) Shut the door! You will not shut the door.

These types of examples have been noticed at least since Black's (1952) paper. On the surface, they seem to have a certain similarity to OMP and CMP, but in their most natural reading they are not (fully) assertions: their first conjunct is a promise (21), a request (22), or a command (23). What would it mean for Priority? Woods (2014; 2018) argues that counting (21)-(23) as Moorean would be detrimental to it since such utterances do not express beliefs and we might think of some other examples of Moorean speech acts that do not express mental states at all. A supporter of Priority is then challenged not only to explain infelicity of Moorean *assertions*, but of all speech acts in "belief" terms. In this chapter I will do so, by extending the Stalnaker's theory of the "essential effect" an assertion has on the common ground to other, non-assertoric speech acts.

1.4.2. Two questions concerning Moorean thought

After finishing the assessment of different problems of Priority, in the second part of the dissertation, I will attempt to track the sources of the irrationality of Moorean belief. If the Priority thesis is true (as I will argue in the first part of the dissertation), then an explanation of the irrationality of Moorean beliefs would suffice to solve Moore's Paradox in general.

The first question I will address concerns the type of irrationality and therefore a source of paradoxicality of Moorean beliefs. I will try to answer it by comparing two different approaches one may have toward such an explanation. The first one gives us a straight answer to the question – irrationality of Moorean beliefs stems from the fact that they are simply impossible to be rationally believed; in Roy Sorensen's useful phrase, they are doxastic or epistemic "blindspots" (1988). This strategy, which I shall

label "introspectionist", crucially postulates that the source of Moorean irrationality lies in a form of "self-blindness" (Shoemaker 1995), i.e., the striking failure to properly introspect one's beliefs; to prove that one cannot believe Moorean conjunctions *simpliciter*, one seemingly needs to appeal to the principle of positive belief introspection, *BB*. Another approach, which tries to avoid such commitments, may be labeled as a "self-defeat strategy". This account sees the irrationality of Moorean beliefs in their "self-falsifying" character, which is thought to be similar to the inconsistency of "pragmatic paradoxes" (O'Connor 1948) such as:

(24) I am not using language now.

(25) No one ever uttered a sentence in English.

The idea is that Moorean sentences, similarly to (24) and (25), may be true, but turn out to be false *when they are asserted or believed*. Their paradoxical nature comes then not from the fact that they cannot be believed, but rather that they cannot be *truly* believed. This strategy has the seemingly appealing feature of not carrying the weight of the strong epistemic principles mentioned later. It is also consistent with a treatment offered by theorists who argue that (analogously to assertion) the norm of rational belief is *knowledge* (Williamson 2000, Adler 2002, Huemer 2007) or *truth* (Shah, Velleman 2005, Williams 2023). I will argue, however, that such a strategy cannot accurately capture all Moorean phenomena – both because of the logical weakness it is committed to, and because of more general philosophical problems. These considerations would result in the conclusion that, after all, an explanation of Moore's Paradox needs to be cashed out in terms of violations of the positive introspection principle – *BB* – along the lines of introspectionist strategy.

In the second chapter of this part, chapter 5., I will proceed to defend this conclusion. Since introspection principles faced a lot of criticism in epistemology and

philosophy of mind, there is a need to present a philosophically neutral justification. It is especially important if the positive introspection principle is expected to serve as an *explanation* of Moore-paradoxicality. Many philosophers who supported this principle, such as Shoemaker (1995) or Hintikka (1962), treated Moore-paradoxicality as *data* supporting its introduction. This, of course, may be treated as a good abductive argument for *BB*; but if it were to genuinely explain the rise of Moore-paradoxicality, we cannot justify it by the fact that Moore-paradoxicality arises. Therefore, we need more general arguments for the plausibility of this principle at least in the realm of beliefs that may give rise to the paradox.

The strategy that I will employ adheres to the tradition of so-called "transparency" accounts of self-knowledge, inspired by Gareth Evans' famous remarks in "The Varieties of Reference" (1982) and represented most profoundly by Tyler Burge (1996), Richard Moran (2001) and Alex Byrne (2005, 2018). The general idea of such approaches may be summarized by the thesis that our introspective abilities are not based on any type of special relation of acquaintance between us and our mental states, but rather grounded in reasoning from external premises. While some (Das, Salow 2018) argued that transparency accounts do support not only *BB* but the even more controversial *KK* principle, I take much of the resistance against *BB* to be fueled by the widespread rejection of the latter principle and the epistemic externalist intuitions guiding it. Therefore, in this last chapter, I will examine the prospects of a moderate solution, that saves *BB* (or more precisely, its stronger counterpart, *KB*) as a natural consequence of transparency accounts of self-knowledge, while rejecting *KK*. In more detail, I will argue that transparency accounts do support the adoption of *BB*, but not *KK*, and that the most popular arguments (in particular, Williamson's 'margin of error' arguments) against *KK* do not threaten *BB*. Therefore, one may defend *BB* as an independently plausible principle guiding conscious beliefs without accepting seemingly problematic positions usually associated with it and provide a universal

and independently grounded account of why Moorean sentences cannot be rationally believed.

In summary, in this dissertation, I propose a two-step approach to Moore's Paradox. In the first part, I defend the Priority thesis and argue that the infelicity of Moorean assertion is to be explained in terms of irrationality in belief. In the second part, I argue for a specific account of irrationality of Moorean beliefs, that is – irrationality stemming from a violation of the principle of positive introspection. If I succeed, this dissertation may be treated as a proposal of a universally applicable solution to Moore's Paradox in speech and thought.

1.5. Methodological remarks

Before moving on, I need to remark on certain methodological questions. Firstly, since in the above discussion, I criticized the "syntactic" characterization of Mooreanness as a specific type of conjunction, I need to provide an alternative, so the proposed solution does not overspill on unrelated phenomena. Secondly, as both in this chapter and later I depend on intuitive judgments concerning the felicity and rationality of performing certain speech acts or believing their content, the question of how such judgments are justified needs to be addressed. Lastly, I will discuss the issue of and whether Moore-paradoxicality ought to be characterized in sentential or propositional terms.

1.5.1. On defining Moore-paradoxicality

As said above, the simplistic way of characterizing Moorean sentences is *via* their similarity to the *syntactic* form of OMP or CMP. In my view, this custom for a long time led people to consider only the assertions of CMP and OMP as something that merits an explanation and, in turn, lean towards pragmatic solutions focused on explaining the absurdity of consecutive assertions of *p* and "I don't believe that *p*" or

"I believe that ~*p*". This focus led many to disregard or not consider sentences such as those mentioned at the end of section 1.2.3.; though such sentences may vary in the instantaneous perceived infelicity of asserting them, this could be plausibly explained by the time required to logically parse them. As they all (logically or semantically) entail OMP or CMP, they should be classified as infelicitous to assert and irrational to believe despite their different syntactic form. One may retort that it is then the *logical form* and *equivalence* we should be concerned with here. In this view, a Moorean sentence would be a sentence whose logical form is equivalent to, or entails OMP or CMP. Such a definition however suffers from the lack of specificity, especially for the used notion of *equivalence* and *entailment*. To see this, one may note that the following two versions of a *commissive* Moorean case:

(CMP) *p*, but I believe that ~*p*.

(NegCMP) ~*p*, but I believe that *p*.

are usually thought to have the same logical form, since (NegCMP) is classically equivalent to (CMP) with ~*p* in place of *p*. This equivalence, however, does not hold in intuitionistic logic, since to demonstrate it, one needs to use the law of double negation (Peluce 2017). Does it mean that the intuitionist actually has not *two* (the omissive and commissive), but *three* (the omissive and two commissive) distinct paradoxical forms? Is there a specific kind of negation that appears in Moorean sentences – classic, intuitionistic or other? In my opinion, we cannot settle these questions without a

principled reason, at least on the stage of simply defining the class of problematic sentences[28,29].

Another problem with the two approaches mentioned above (syntactic or logical characteristics) is that they either rule out the intuitively related phenomena, such as versions of Moorean sentences using other verbs than "believe", such as EOMP, COMP, or require arbitrarily treating them as additional "Moorean" fixed points of logical or syntactic equivalence. Where does the intuition that such sentences are Moorean come from? It seems that it is suggested, again, by syntactic similarity with OMP. But is such similarity sufficient (even if, as I argued above, not necessary)? Consider also the following sentences (in (26), interpret "might" in its epistemic reading):

(26) *p*, but it might be that ~*p*.

(27) *p*, but it's not certain that *p*.

Are such sentences Moorean or not? They are syntactically similar and have an air of absurdity to them, and they are not, at the face of it, self-contradictory – for to say that something is not certain or possibly false does not entail that it's false. It may

---

[28] One should also note that such logical equivalence should be defined in propositional, not sentential terms – as I demonstrate in 1.5.3., defining Moore's Paradox in such way is problematic in itself.

[29] In the literature, one may also encounter definitions such as that Moorean sentences are sentences that are "doxastically self-defeating" (Williams 2013a,b, Rieger 2015). Such definitions are however not theoretically neutral, for they presuppose the completeness of a certain explanatory strategy. What I seek here is a strategy-independent characterization – to claim that a strategy successfully explains Moore's Paradox one cannot simply start by saying that Moorean sentences are only those susceptible to strategy's treatment. I will come back to this issue in Chapter 4.

seem that therefore we should include them on the same basis as EOMP, JOMP, or COMP. But as Seth Yalcin (2007) notes, they profoundly differ from them and such sentences containing epistemic modals "give rise to their own sort of 'paradox', one that differs from Moore's paradox in significant respects. (…) It turns out these conjunctions are much more difficult to felicitously embed than Moore-paradoxical sentences, and careful attention to this fact points to some interesting constraints on any theory of the meaning of epistemic modal operators" (Yalcin 2007, p. 985). To demonstrate this, Yalcin asks us to consider the following:

(28) Suppose that: *p*, but it's not certain that *p*.

(29) Suppose that: *p*, but I don't believe that *p*.

It seems clear that while (29) is unproblematic, (28) may still be judged as infelicitous. This is why Yalcin and others (e.g., Mandelkern 2019) distinguish (26)--like constructions containing epistemic modals from Moorean sentences and label them "epistemic contradictions" or "Wittgenstein sentences".

Simultaneous consistency and irrationality of asserting or believing, phenomenologically perceived "feeling of Moorean absurdity", syntactic similarity or logical equivalence with paradigmatic cases are all not sufficient (and, I would argue, apart from the first condition also not necessary) for a sentence to be Moorean. This, of course, does not mean that none of them can guide us in finding examples; but they are, at best, heuristics not constraints. What then may we propose instead? In my view, there is little hope for a proper definition of Moorean sentences without making some theoretically objectionable choices or presupposing the success of some explanatory strategy. But this does not mean that we cannot at least try to operationalize the notion of a Moorean sentence and propose some tests that establish whether a sentence is Moorean or not.

As we saw above, the earliest puzzling characteristics of Moorean assertions apart from their simultaneous consistency and unassertability, were their first/third person asymmetry, that is the fact that while "*p*, but I don't believe that p" sounds absurd, "*p*, but she doesn't believe that p" does not, and analogous past/present tense asymmetry[30]. Another characteristic observed in Wittgenstein's letter to Moore was that such sentences may be felicitously supposed – "[i]t makes sense to say "Let's suppose: p is the case and I don't believe that p is the case", whereas it makes no sense to assert "p is the case and I don't believe that p is the case"" (McGuiness 2008, p. 365). We may note here, that Yalcin's argument mentioned above implicitly makes use of this Wittgensteinian condition as a way of distinguishing Moorean sentences from his "epistemic contradictions". The first two asymmetries point out also that Moorean sentences need to be perspective sensitive, in a sense in which they need to be subject to changes in grammatical person and tense. This also crudely aligns with the exclusion of "epistemic contradictions", for while (26) and (27) may be felicitously uttered in the past tense, they are not subject to changes in grammatical person.

We may then propose the following set of conditions. Provided that *s* is consistent, *s* is a Moorean sentence if:

(THIRD) It is typically irrational or infelicitous to assert/believe/… *s* in the first-person present tense, but not in the third-person present tense.

(PAST) It is typically irrational or infelicitous to assert/believe/… *s* in the first-person present tense, but not in the first-person past tense.

---

[30] These are the most standardly used cases of grammatical contrast; one could also argue that first-person future tense Moorean sentences (e.g., "*p*, but I *will* believe that ~*p*") are not typically infelicitous or "absurd" (see Williams 2006, p. 237).

(SUPPOSE) It is typically irrational or infelicitous to assert/believe/… *s* in the first-person present tense, but not to suppose it.

These conditions, in my opinion, should not be read as a definition of "being Moorean", but rather as a strong indication of it. If a sentence satisfies all conditions, I treat it as a strong argument that it should be treated as Moorean and fall under the scope of any proposed explanation. If a sentence fails some (as Yalcin's epistemic contradiction) or all three tests, it may be used as an argument for it not being Moorean. Moorean speech acts and mental states will be characterized as those exhibiting this sort of asymmetry with respect to some Moorean sentence.

The conditions mentioned above, although incomplete, allow us to set the scope of what is to be explained by a unified strategy of explanation. In the following dissertation, I will use them to decide whether a specific case of a speech act or a mental state should be regarded as Moore-paradoxical and therefore – whether it ought to be connected to the possession of irrational beliefs, if Priority is true.

1.5.2. Intuitive infelicity and irrationality judgments

The above tests, which I take to be primary guides in identifying Moorean sentences, put stress on the classification of certain beliefs as irrational and speech acts as infelicitous; moreover, in this chapter, I have relied mostly on intuitive judgments supporting such classifications. A curious critic could perhaps ask what justification exactly I, or any other philosopher or linguist, have for these judgments. This is perhaps one of the most uncomfortable questions to ask a philosopher of language or epistemologist, for such judgments are supposed to be "obvious" data one builds their theory on. But how these data are actually obtained?

Due to its lineage coming from late Moore and Wittgenstein, it is not only apt to call Moore's Paradox "a child of analytic philosophy", but even of ordinary language philosophy specifically. After all, it is ordinary language philosophers' attention that led to giving Moore's observation such philosophical prominence; and their keen focus on the actual *use* of expressions gives rise to the whole methodology of classifying certain speech acts as "felicitous" or "infelicitous". But since its beginning their classifications had been described as tendentious and authoritative, as put into words by a fierce critic of ordinary language philosophy, Herman Tennessen:

"Various devices have been proposed for distinguishing impermissible from permissible or legitimate locutions. (...) The easiest way by far is simply to "*see*", to "*hear*", or otherwise *perceive* the permissibility or the impermissibility of a given locution. A native speaker, it has been maintained, will never (or rarely) be in doubt. He perceives the (im-)permissibility directly, instantaneously, in a flash of revelation, by some sort of linguistic instinct, logical sense, hermeneutical clairvoyance. Oddly enough, these seem to be qualities most often found in precocious children, and pedantic logicians." (Tennessen 1961, pp. 266-267)

This quite vicious characterization of ordinary language philosophers has, nevertheless, a ring of truth to it, as some speech act's infelicity, labeled with "#" (or, in more controversial cases, "??"), often seem to be based on nothing else than just author's opinion. Though it might seem that Moorean sentences' infelicity status is undisputable, it was one of Tennessen's targets in the quoted article. In his study (1959, also reported in the abovementioned article), Tennessen found that when asked to judge Moorean sentences, people tended to describe them as "self-contradictory", when prompted with a "logico-maniacal" lecture on the importance of precise and logical speech. When taking a questionnaire was, however, preceded by a "common-sensical" lecture on the importance of understanding speaker intentions, practically

none of the subjects classified Moorean sentences similarly (1959, pp. 377-380), instead interpreting them as akin to "I can hardly believe that *p*"[31].

Of course, the results obtained by Tennessen could be provided with a natural charity-based interpretation – when his subjects encountered a seemingly "contradictory" statement, they reinterpreted it in a way that conveyed some non-literal meaning which could accord with the cooperative speaker's intentions (as it canonically happens when the speaker says something obviously false or obviously true). His observation does not mean that Moore's Paradox is, somehow, dissolved or could be described merely as ordinary language philosopher's fable – the fact that consistent and possibly true sentences *can* elicit judgments of "self-contradiction" is in itself still puzzling, and that is what the first part of Tennessen's study shows[32]. Yet, this insight still calls into question the method we rely on when we go on to classify Moorean sentences or their variants (such as EOMP, COMP, or JOMP): are they "typically" infelicitous to assert or irrational to believe, or do such judgments require ill will or peculiar judgment? Even more than judgments of ungrammaticality[33] (standardly marked with "*" instead of "#"), judgments of pragmatic felicity seem to be "graded" and contextually varied[34] (Ariel 2010, pp. 42-45), and the similar seems to

---

[31] Sorensen also notes that Moorean sentences "have common currency as expressions of surprise" (2007, p. 38).

[32] One may find further (though scattered) evidence from comparative and developmental linguistics that Moorean assertions (and other Moorean speech acts) are *more often than not* classified this way by ordinary speakers (for some examples in the respective fields, see Faller 2002, Bartha 2021). I have not, however, encountered any systematic psycholinguistic investigation of felicity or acceptability judgments regarding Moorean sentences.

[33] For an extensive discussion of methods used to justify judgments of ungrammaticality, as well as some skeptical doubts, see Schütze 2016.

[34] Though it is still not always trivial to differentiate between the two; see Bar-Hillel 1971.

be true of irrationality judgments[35]; it is natural to expect some interpersonal variance in forming them, as well as certain cases in which forming such judgments would be challenging and their truth far from secure. One also need not look far to find contexts in which assertions of OMP or CMP are judged by many to be felicitous, even if they are interpreted literally (or at least closer to its literal meaning). Eliminative materialist's (of Patricia Churchland type) assertion: "*p*, but I don't believe that *p* (since there are no beliefs)" does not elicit similar disapproval as "standard" cases of OMP assertion (Turri 2010)[36]. The obvious and often-employed solution is to classify such cases as "atypical"; as noted in 1.1, Moorean observation remains paradoxical even if we limit our attention to "typical circumstances". But when are we justified in calling certain expressions or thoughts *typically* infelicitous or irrational? This question cannot be plausibly resolved here without proper empirical investigation.

Due to the subject matter of this dissertation, relying on judgments concerning "typical" infelicity of speech acts or irrationality of beliefs is unavoidable and (for practical reasons) these judgments need to be based on intuitive assessment. I wish nevertheless to remain cautious about them. In the first two chapters concerning linguistic matters, I will abstain from using the #/?/?? signs, as I think they would signify unwarranted certainty in linguistic judgment; I will treat examples of intuitive irrationality judgments similarly. Instead, I will provide numbered examples of sentences and describe them as "intuitively felicitous" or "infelicitous" only in the

---

[35] The issue concerning the legitimacy of "irrationality" judgments and ascriptions is, as well, understudied, though relying on an intuitive assessment is common. Their source seems to, similarly to felicity judgments, lie at large in the use of a plurality of similarity-based heuristics. See Bortolotti 2004 for some elaboration in the context of delusional beliefs.

[36] A survey of many similar cases can be found in Hájek 2007. For a proposal on how to delineate them from Moore-paradoxical assertions and beliefs *proper* see Coliva 2015; for my analysis of similar constructions in religious discourse, see Tarnowski 2023.

body of the text, allowing the reader to confront their intuitions with mine on the matter; same goes for "irrationality" ascriptions. Whenever possible, I shall explicitly cite the source of such intuitive judgments, deferring to others who decided to report them in print. Otherwise, the reader should treat them as expressing intuitions of the author that he thinks have a chance of being universal and hope the reader shares.

1.5.3. Propositions and sentences

The last thing that needs to be addressed is the question of why throughout this chapter I spoke about Moorean *sentences*, not *propositions*. As it is quite widespread in epistemology (but, curiously, not in the philosophy of language) to define Moore-paradoxicality in propositional, not sentential terms (cf. Sorensen 1988, DeAlmeida 2001), this choice may strike a reader as peculiar. One could even see speaking of "believing" a *sentence* to be a category mistake, given that this verb is standardly followed by a propositional "that"-clause. This usage may be, of course, treated as a convenient shorthand for "believing that sentence *s* (or its accurate translation) is true"; one may also point out many influential sentential analyses of belief, e.g., in terms of affirmative disposition towards object sentences (Carnap 1988, pp. 53-55) or in terms of "tokening" a sentence of Mentalese (Fodor 2008, p. 68). However, given the relative unpopularity of such approaches, I need to justify my choice. In this subsection, I will shortly elaborate on this decision and its certain qualifications.

The main reason for taking the sentential route lies in the fact that while there is no substantive dispute as to what sentences are and how they ought to be individuated (*modulo* the intricacies of metaphysics of linguistics), the same is not true of propositions. Even granting that metaphysical disputes regarding the status of *abstracta* and whether propositions are structured or unstructured entities could be perhaps put aside, we cannot ignore the obvious question of the coarse- or fine-graininess of content. Let me demonstrate how such issues impact the prospects of characterizing Moore-paradoxicality in propositional terms. As a starting point,

consider the coarse-grained view according to which propositions, understood both as contents of belief and assertion, ought to be identified with unstructured sets of possible worlds or functions from possible worlds to truth values (Stalnaker 1984), or a more fine-grained neo-Russellian view (e.g., Salmon 1986), according to which singular propositions contain directly the object they are about. Both of these views, if one assumes Millian analysis of proper names, lead to the conclusion that the following sentences express the same proposition:

(30) Hesperus is a planet.

(31) Phosphorus is a planet.

If that's however the case, all of the following sentences should express either commissive or omissive Moorean proposition:

(32) Hesperus is a planet, but I believe that Hesperus is not a planet.

(33) Hesperus is a planet, but I believe that Phosphorus is not a planet.

(34) Hesperus is a planet, but I don't believe that Hesperus is a planet.

(35) Hesperus is a planet, but I don't believe that Phosphorus is a planet.

I take it as intuitively true that while asserting or believing (31) and (33) is genuinely puzzling, the same does not apply to (32) and (34), at least if an agent believes "Hesperus" and "Phosphorus" to name different entities (Venus and some star, for example). Similar problems emerge if one, following Kaplan (1989), takes indexicals and demonstratives to be directly referential. Borrowing the case from Perry (1979), one could imagine a case in which, unbeknownst to me, I am observing a reflection of a person in the mirror without realizing that it is myself; knowing that the person's pants are on fire, the following could have crossed my mind:

(36) My pants are not on fire, but I believe that his pants are on fire.

(37) My pants are not on fire, but I don't believe that his pants are not on fire.

Both (35) and (36), though not Moorean according to the definition in 1.5.1, express the same proposition as their counterparts in which demonstrative determiner "his" is replaced with "my", given coarse-grained theories (for elaboration on this point see Chan 2010, Williams 2023). This means, implausibly, that they *should* count as Moorean, if Moorean assertion or belief is characterized only in terms of the relation between the speaker/agent and coarse-grained propositional content.

One response to similar problems would be to retort to a more fine-grained theory, under which (31)-(32) and (33)-(34) shall be classified as expressing different propositions, as well as (35)-(36) and their uniformly first-person counterparts. Such proposals could include some interpretations of classical Fregean theory, certain neo-Fregean theories (e.g. Zalta 1988), as well as some proposals that incorporate the sentential syntactic and lexical structure criteria in the principles of individuation for propositions (for early precursor, see Ajdukiewicz 1967; for contemporary sources see, e.g., King, Soames, Speaks 2014). On such approaches, Moorean propositions could be characterized perhaps as straightforward abstract analogs of Moorean sentences.

A defender of the coarse-grained approach to propositions may instead respond by arguing that belief (or assertion) should not be fully characterized as a *binary* agent-proposition relation, but a ternary one, predicated of the agent, the proposition, and the *guise* under which the proposition is believed or asserted[37]. Just as on the other side

[37] This echoes Kaplan's (1989, pp. 529-540) suggestion that while the content of indexical belief should be characterized in coarse-grained propositional terms, their "cognitive significance" is to be identified with its mode of presentation.

of the debate, there is a plurality of different theories of guises, ranging from those treating them as individually possessed notions or ideas (Crimmins, Perry 1989) to linguistic approaches, according to which "to believe a proposition under a guise *S* is to believe the proposition p expressed in context by *S* and to take *S* as an articulation of one's belief that p" (Hong 2021, pp. 1882-1883). In such theories, Moorean propositions would then be irrational to believe insofar as the act of believing them and the belief (or lack of belief) predicated of oneself contained in the propositional content were both under a uniform guise.

It is quite clear that a pure characterization of Moorean propositions would require making careful choices concerning one's preferred theory of propositional content and metaphysics of belief. But more often than not, these theories would approximate a sentential reading, either by characterizing Moorean belief or assertion as a belief in or an assertion of a Moorean proposition under a certain *sentential* guise or by settling on a picture of propositional content under which propositions are almost as fine-grained as sentences. As I wish to remain as non-committal as possible with respect to the choice of a theory of propositional content, I prefer to characterize Moore-paradoxicality in sentential terms and to leave it to the reader's theoretical preferences to decide whether they wish to read "believing" or "asserting" a Moorean sentence *s* as either "believing" or "asserting" a proposition under *s*-guise or a fine-grained proposition that expresses *s*'s content, or simply adopting a sentential analysis of belief and assertion alike to those mentioned above[38,39].

---

[38] Another obvious perk of speaking in sentential terms comes from the fact that in Chapter 3. I will discuss not only Moorean assertions and beliefs but also other Moorean speech acts, such as promises or orders, which are easier to characterize in sentential than propositional terms.

[39] I should note in the end that not all problems of disguise can be *outright* dealt with by resorting to sentential terms. Kripke's (1979) Peter, for example, might sincerely assent to the Moorean sentence: "Paderewski is a great musician, but I don't believe that Paderewski is a great musician", thinking

The problems discussed here are all well-recognized and their treatment is the topic of extensive literature, extending far beyond the scope of the present dissertation. The point of bringing them up is, first and foremost, to not fall into them later on and pre-emptively settle on what I take to be the most neutral characterization of Moore-paradoxicality which avoids the problems connected with them. In the next chapters, these issues will not be further brought up. In chapter 3., for terminological convenience and to maintain continuity with the existent literature, I will use the characterization of "conversational common ground" in propositional terms: the qualifications mentioned in the above paragraph regarding the interpretation of propositions as objects of belief, knowledge, and assertion will apply. In chapters 4. and 5. I will also utilize normal epistemic and doxastic modal logics and discuss the philosophical justification of their axioms. Their widely adopted Kripke-style semantics, if given philosophical consideration, usually leads to analogous troubles. If the reader wishes, I see no explicit reason for interpreting the results discussed there in syntactic terms (in the style of van Wright [1951]) or translating them to non-modal systems[40], and treating semantic proofs as serving merely better exposition. Nothing of crucial importance depends on it.

---

wrongly that the first and second token of "Paderewski" are not co-extensive; we might also (following, e.g., Crimmins 1992) think of similar cases involving demonstrative reference. One can, however, defend a view (as I did in Tarnowski, Głowacki 2022) that these problems call rather for a more fine-grained view of what constituents of sentences (e.g. proper name or demonstrative types), rather than of propositions, are.

[40] In line, for example, with Marciszewski 1972 or Tokarz 1990, 1993. It is telling that such attempts usually define "belief" in sentential terms analogous to Carnap's (see: Łoś 1948, p. 70; Marciszewski 1972, p. 98; Tokarz 1993, pp. 157-158); though see Rescher 1968, pp. 49-52 for dissent. For an extensive overview of different similar proposals and systems, see Lechniak 2011.

1.6. Conclusion

This chapter served a fourfold role. Firstly, section 1.1, introduced Moore's Paradox as both an epistemological and linguistic puzzle taking the form of the following argument from intuitively plausible premises to an inconsistent conclusion:

MOORE'S PARADOX:

(a) In typical circumstances, *p* may be rationally believed and felicitously asserted if it is true.
(b) Moorean sentences (e.g., OMP and CMP) may be true.
(c) [from (a)-(c)] In typical circumstances, if Moorean sentences are true, they may be rationally believed and felicitously asserted.
(d) In typical circumstances, Moorean sentences cannot be rationally believed and felicitously asserted, even if they are true.

As noted, the plausible source of this puzzle, given the accuracy of Moore's observation expressed by (d), is the intuitive principle expressed by (a). The sound method of resolving Moore's Paradox is therefore to propose an alternative constraint on rational belief and felicitous assertion than the truth of its object sentence, provide its plausible philosophical justification, and demonstrate, how it allows to show irrationality of believing or asserting Moorean sentences.

Secondly, this chapter introduced and assessed the historical context of the study of the paradox. In section 1.2, I discussed two earliest attempts at solving it: Moore's and Wittgenstein's, and more contemporary (neo-Gricean and expressivist) approaches inspired by them. Through subsections 1.2.2 and 1.2.3 I argued that these attempts were unsuccessful and hypothesized (in subsection 1.2.4) that their failures resulted from methodological neglect of the doxastic dimension of the paradox and

focusing only on its linguistic side. This led me (in section 1.3) to postulate that, to get a satisfying and unified solution for the paradox, one needs to start with an explanation of the irrationality of Moorean beliefs and supplement it with a theory connecting the irrationality of a belief with an infelicity of an assertion. Following Shoemaker (1995), I stated the central hypothesis of this dissertation as follows:

(**Priority of Belief**) The infelicity of Moore-paradoxical assertions is explained by the irrationality of Moore-paradoxical beliefs.

and hypothesized that the plausible link between belief and assertion takes the form of the following "bridging principle":

(**Bridging Principle**) One can warrantably assert only what one can rationally believe.

In section 1.4. I outlined the plan of the dissertation to follow, which aims to fill the blanks in the proposed explanation. The plan breaks into two supplementary parts concerning the explanation of Moore-paradoxicality in speech and thought, both consisting of two chapters. In Chapter 2., I will examine the normative accounts of assertion present in the literature to find the most plausible theory that both justifies the Bridging Principle and explains the felicity data concerning the assertions of various forms of Moorean sentences. In Chapter 3., I will extend the proposed explanation to cover non-assertoric speech acts claimed to exhibit infelicity similar to Moorean assertions. In the second part, devoted to Moore-paradoxicality in thought, I will first assess (chapter 4.) the comparative strength of two proposals for explaining the irrationality of Moorean belief: "self-defeat" and "introspective" strategies. Since

the outcome of this comparison will lead me to accept the accuracy of an introspective account, in Chapter 5. I will defend its most philosophically contentious part: the positive introspection principle for belief. As the most popular arguments against it tend to rely on epistemic externalist intuitions implicit in Timothy Williamson's attacks on an analogous principle for knowledge (*KK*), I will demonstrate that neither externalism nor rejection of *KK* entails rejection of similar principles for belief.

The fourth and last purpose of this chapter was to discuss methodological issues connected with the study of Moore's Paradox adopted further in the dissertation, which was done in section 1.5. In the first subsection 1.5.1, I argued that while it is impossible to give a theoretically neutral *definition* of Moore-paradoxicality, one may adopt the following list of sufficient conditions (or "tests") for it (with *s* being a consistent sentence):

(THIRD) It is typically irrational or infelicitous to assert/believe/… *s* in the first-person present tense, but not in the third-person present tense.

(PAST) It is typically irrational or infelicitous to assert/believe/… *s* in the first-person present tense, but not in the first-person past tense.

(SUPPOSE) It is typically irrational or infelicitous to assert/believe/… *s* in the first-person present tense, but not to suppose it.

This definition gives rise to two further methodological questions that were clarified in subsequent sections 1.5.2 and 1.5.3. As these tests employ intuitive rather than principled judgments of typical "infelicity" and "irrationality" of speech acts and thoughts, it is important to realize that they are contentious, which led me to introduce various qualifications and methodological choices concerning the use of such

"intuitive" data. The second question concerned the issue of defining Moore-paradoxicality in sentential, rather than propositional terms. Though the latter characterization is more popular, it is problematic unless one settles on one or the other highly contentious theory of propositional content or semantics of belief ascription. Because resolving this hotly debated issue cannot be given justice in this thesis, I decided to employ sentential terminology, while at the same time providing few possibilities for translating it to propositional terms without adjudicating between them.

What follows from now on is simply the execution of the plan announced here. Though Chapters 2-5. had been originally written as standalone pieces later adjusted to form a full monograph, I hope that this introduction makes it clear that together they stand as a complete attempt at solving Moore's Paradox in speech and thought.

Part I: Moorean Speech

Speaking broadly, verbal habits crystallise our beliefs, and afford the most convenient way of making them explicit. To say more for words is to fall into that superstitious reverence for them which has been the bane of philosophy throughout its history.

(Russell 1926, p. 642)

Chapter 2. Moore's Paradox and the Norms of Assertion

Section 1.3. of the previous chapter proposed the following principle to serve as a theoretical bridge between the rationality of belief and warrantability of assertion:

(**Bridging Principle**) One can warrantably assert only what one can rationally believe.

As explained, this principle allows us to describe the infelicity of Moorean assertion along the lines of the Priority thesis:

(**Priority of Belief**) The infelicity of Moore-paradoxical assertions is explained by the irrationality of Moore-paradoxical beliefs.

provided that one endorses the plausible claim that the assertion's unwarrantability leads to it being perceived as infelicitous by ordinary speakers. The Bridging Principle needs philosophical justification; even more importantly, it needs to be shown how it handles a range of linguistic data concerning the infelicity of assertions other than those of the classic CMP and OMP forms. In the first chapter, I briefly pointed out that this work can be done by choosing an appropriate normative account of assertion, that is – a characterization of the speech act of assertion that portrays it as being subject to normative, epistemic constraints. In this chapter, I undertake this task in more detail and evaluate different norms of assertion defended in the literature with respect to their consistency with the Bridging Principle and the ability to explain a variety of forms Moorean assertions take.

Despite the semi-reviewal nature of the first part of this chapter, I shall focus extensively only on what had been dubbed in the literature (see Pagin and Marsili 2021) as "content-oriented" norms that emphasize the epistemic relation between the speaker and the content of their assertion. Therefore, I shall not discuss in full justice, e.g., analyses of assertion in terms of the speaker's commitments (e.g., Brandom 1983). However, I will demonstrate, whenever possible, how the problems or strengths of a given content-oriented analysis may be extended to a class of other approaches sharing important similarities. Within this class of norms, I shall distinguish doxastic and non-doxastic norms (the former demanding that the speaker needs to believe the content of their assertion and the latter not imposing such condition) and weak and strong doxastic norms (differentiated by the demand of one's belief in assertions content to be justified).

In short, this chapter will conclude that a proper explanation of Moorean infelicity in speech needs to be accounted for using a discourse-relative strong doxastic norm of assertion. In section 2.1. I will briefly comment on the difference between normative and descriptive accounts of assertion and the "conventional expression of belief" descriptive condition. In section 2.2., I shall argue that only strong doxastic norms can be regarded as continuous with an explanation of Moore-paradoxicality of beliefs in a way demanded by Priority and that the available alternatives (non-doxastic or weak doxastic norms) suffer from independent theoretical drawbacks. In section 2.3 I will discuss such strong doxastic norms, which are usually seen as providing the most convincing explanation of the unassertability of Moorean constructions. I will argue that the picture in question needs to be more complex to capture the context-dependent differences in felicity judgments regarding Moorean assertions; in turn, in section 2.4, I shall propose a general framework for relativizing a norm of assertion with respect to the strength or type of justification required for a warranted assertion given the specific discourse.

2.1. Normative and descriptive accounts of assertion

Before delving into comparing different proposals, we need to quickly clarify the meaning in which we speak of "normative accounts of assertion" as opposed to "descriptive accounts". In short, while the former describe what specific norms the assertor is expected to follow, the latter aim at telling us what an assertion *is*, that is – what distinguishes it both from entirely different speech acts such as promises or commands, as well as other constatives, such as conjectures or hypotheses. Just as there is a plurality of different norms proposed for assertion, there is abundant literature containing different descriptive characteristics of assertion in terms. As noted in (Pagin, Marsili 2021), many of such descriptive endeavors can be derived from two interpretations of Frege's definition of assertion as "the manifestation of (...) judgment" (1918/1956, p. 294), where "judgment" either stands for an acknowledgment of the truth of the object proposition (Frege's *Thought*) or simply for a belief. Assertion is, in turn, often pictured as either an act of "presenting as true" or "signifying belief"; commonly, these two analytic strands were also supplemented by analyses of the role assertions play in communication often inspired by Gricean pragmatics (e.g., Stalnaker 1978, Bach and Harnish 1979).

As I indicated in 1.2 and 1.3, relying on a purely descriptive understanding of assertion or its standard pragmatic/psychological effects would not allow us to explain the infelicity judgments concerning Moorean assertions (some of them would be expanded upon in 2.2.2). *Prima facie* such judgments signify a normative breach: in asserting a Moorean sentence, the speaker ostensibly violates our expectations of what one *may* assert, as this sort of pragmatic infelicity cannot be explained by apparent ungrammaticality, nonsensicality, or obvious falsity of their utterance[41]. Nevertheless,

[41] The norms I discuss here are in a broad sense, norms of epistemic, not moral or prudential propriety.

to classify some utterances of Moorean sentences as *assertions* in the first place, we need certain descriptive resources. For now, I shall adopt the guiding idea of Bach and Harnish (1979) that types of speech acts should be individuated based on the type of attitude they conventionally express, and take assertion of *p* to conventionally express the belief that *p*. By conventional expression of belief, I will mean, after those authors, constituting a (defeasible) reason for the hearer to ascribe a belief that *p* one the basis of one's utterance of "*p*" in the declarative mood; though I think that one may quite intuitively grasp the intended meaning and modify this definition. The infelicity judgments of our interest will, in turn, be based on evaluating whether such conventional expressions, already recognized as acts of assertion, meet relevant normative conditions of propriety.

This two-stage approach is not the only option; some philosophers argued that a normative account of the assertion may serve also a descriptive purpose, resulting in a simpler and stronger theory. Timothy Williamson (2000, pp. 239-242), who defends the knowledge norm of assertion (see section 2.3), famously proposed that it should be understood as a *constitutive* rule: not only regulating the speech act of assertion but also essentially specifying what it means to assert. This proposal is, however, very controversial (see, e.g., Cappelen 2011[42], Maitra 2011, Pagin 2016) and often depends on what we take the "constitutive rule" to mean[43]. Clarifying is however problematic,

---

[42] Instead of insisting on a primacy of the descriptive route of defining assertion, Cappelen defends a radical view according to which "assertion" as a whole category is hopelessly defective. The criticisms presented by him in (2011) are, however, almost exclusively focused on Williamson's "constitutive rule" view. To explain the infelicity of Moorean "sayings", as he prefers to call them, he follows a neo-Gricean approach which was rejected in 1.2.2.; therefore, I choose to interpret his remarks as aiming at refuting the *constitutive norm* view.

[43] It cannot be meant in the standard way proposed by Rawls (1955) and Searle (1970) as defining when an individual action *counts as* a certain type of action: many utterances that don't meet Williamson's "knowledge" requirement (e.g. lies) would still count as assertions.

especially since Williamson does not offer any positive detailed account of what he takes constitutive rules to be (Williamson 2000, p. 240; for a heavily qualified defense of Williamson see Simion, Kelp 2020; for further discussion see the supplement to Pagin, Marsili 2021). Resolving the issue is, in my opinion, not necessary for the present purposes. Unless only breaking the *constitutive norm*, but not any other, gives rise to infelicity judgments, I need not defend such a strong stance.

In what follows, I will therefore adopt a more traditional view that the speech act of assertion of *p* can be characterized descriptively as a conventional expression of the belief that *p*, while simultaneously being regulated by other norms that play a role of guiding our expectancies towards the assertor. The normative proposals discussed here should then be read as hypothetical sufficient conditions for an assertion to be warranted, without additional baggage of being "constitutive". This issue shall not play a central role in this chapter, though it would later benefit the simplicity of analyses in chapter 3.

2.2. Preliminary Requirements: Justification and Belief

Let me now set some preliminary requirements for a normative account of assertion to be compatible with Priority. The Bridging Principle, our first step towards this compatibility, can be broken into two distinct constraints it imposes on warranted assertion:

(B) Speaker *S* may assert *p* only if *S* believes *p*.

(J) Speaker *S* may assert *p* only if *S*'s justification for *p* suffices for *S*'s belief that *p* to be rational.

In short, B states that the relevant norm needs to be doxastic and J – that it needs to be strong (at least *as strong* as normative demands on belief's rationality). In the

following subsection, I shall defend why both of these conditions are necessary in the explanation of Moorean infelicity data. Though most of the existent normative accounts would agree with B and J conditions, there are exceptions, of which the most prominent are the "Belief" norm (BNA), Lackey's (2007) "Reasonable to Believe" norm (RTBNA), Douven's (2006) "Rational Credibility" norm (RCNA), and Weiner's "Truth" norm (2005). Before discussing which of the norms of assertion meeting B and J criteria account best for the Moorean data, I will first evaluate these proposals and how their authors propose to deal with Moore-paradoxical assertions.

2.2.1. "Justification without belief" constraints

The first type of conflict between J and B conditions with the norms of assertion defended in the literature is the conflict with the class of norms which I dub "non-doxastic" that is – norms that allow assertion to be warranted without imposing any conditions on the speaker, such as the truth norm:

(TNA) *S* may assert *p* only if *p* is true (Weiner 2005).

or norms that only require the speaker to have an appropriate justification for their claim, in the form almost parallelling the J condition[44]:

[44] An interesting case of (strictly speaking) non-doxastic norms is Willard-Kyle's "being in a position to know" norm (2020) and Mandelkern and Dorst's (2022) and Dinges's (2023) "pretending to be certain" norms. These accounts do not appeal to the Gricean strategy outlined below, and they utilize the resources of knowledge norms to get the same verdict on Moorean assertions. I think that the considerations underlying these proposals are quite subtle and, in large part, need not be directly addressed here; as both accounts are committed to the same type of explanation of the Moorean data as

(RCNA) *S* may assert *p* only if *p* is rationally credible to *S*. (Douven 2006).

(RTBNA) *S* may assert *p* only if it is reasonable for *S* to believe *p* (Lackey 2007).

Though I will not go into details of these theories, the same seems to eventually hold for a cluster of theories that characterize an act of assertion in terms of the speaker's commitments without at the same time requiring the speaker to be committed to believing the content asserted[45] – e.g., Brandom's (1983) analysis in terms of dialectical commitment of *being able to respond to legitimate challenges*, Alston's (2000) account in terms of *social liability,* or MacFarlane's (unpublished manuscript) theory that emphasizes the conditions of retraction – as neither of them makes it mandatory that the speaker believes what they assert.

Do these theories allow for warranted Moorean assertions? Obviously, TNA does – the whole point of Moore's Paradox is that Moorean sentences, despite their apparent unassertability, may be *true*; in fact, TNA closely resembles the assertoric part of the (a) premise in the formulation of Moore's Paradox in 1.1 of the previous chapter. Does the same hold for RCNA, RTBNA, and commitment views based on justification? While it might be odd to find Moorean assertions justified or rationally credible, note that we often seem to have perfectly good justification available for certain

---

the Knowledge or Certainty Norms, I shall address them indirectly when discussing these accounts in the next section. Mandelkern and Dorst's proposal will be also indirectly discussed in Chapter 3.

[45] C*ommitment-based views* that do take assertion to necessarily express belief is, e.g., Green's (2020). In what follows I shall regard such views as entailing the truth of doxastic justification norms, without the additional baggage of proving that such norms are *constitutive*. Similarly, I take Brandom's *commitment-based* account to entail RTBNA., while Alston's (2000) TNA.

propositions that we do not believe or even disbelieve for non-epistemic reasons. One may have been presented and even accept the evidence for a certain scientific claim or stance (say, Darwinian evolutionary theory or quantum indeterminacy) and yet disbelieve it because of prudential, religious, or philosophical reasons (one may consider here a devoted creationist or a hardline determinist). Lackey famously presents a battery of such cases in which assertions violating the B condition – which she labels "selfless" – seem to be possible without associated infelicity judgment. Here is one of them, cited *verbatim* from (Lackey 2007, p. 599):

CREATIONIST TEACHER: *Stella is a devoutly Christian fourth-grade teacher, and her religious beliefs are grounded in a deep faith that she has had since she was a very young child. Part of this faith includes a belief in the truth of creationism and, accordingly, a belief in the falsity of evolutionary theory. Despite this, Stella fully recognizes that there is an overwhelming amount of scientific evidence against both of these beliefs. Indeed, she readily admits that she is not basing her own commitment to creationism on evidence at all but, rather, on the personal faith that she has in an all-powerful Creator. Because of this, Stella does not think that religion is something that she should impose on those around her, and this is especially true with respect to her fourth-grade students. Instead, she regards her duty as a teacher to include presenting material that is best supported by the available evidence, which clearly includes the truth of evolutionary theory. As a result, while presenting her biology lesson today, Stella asserts to her students, "Modern-day Homo sapiens evolved from Homo erectus," though she herself neither believes nor knows this proposition.*

Lackey argues that in the CREATIONIST TEACHER (and a range of similar scenarios), the speaker may indeed warrantably assert something they do not believe. But if she is right, then Stella, if she possesses at least some introspective abilities, seems to be able to also say:

(1) Modern-day *Homo sapiens* evolved from *Homo erectus*, but I believe that it did not.

and warrantably assert a commissive Moorean proposition by the lights of RCNA and RTBNA, since the first conjunct will be supported by scientific and the second by correct introspective evidence.

This issue, of course, did not go unnoticed. Both Lackey and Weiner note that the norms they propose seem to make room for the warrantable assertion of Moorean sentences, as they may be true and/or justified; they are not, however, willing to embrace this fact as a consequence of their accounts. Both do, in turn, appeal to broadly neo-Gricean strategies of explanation (Weiner 2005, pp. 237-238; Lackey 2007, pp. 613-615). Lackey and Weiner suggest that what we are dealing with in cases of Moorean assertions is not a clear violation of the norm of assertion, but rather a violation of conversational principles. Weiner argues in defense of TNA that a Moorean assertion fails to meet this condition, as in most circumstances we would take the speaker's unqualified assertion of *p* to be associated with evidence for the truth of *p*, and hence knowledge of *p* (Weiner 2005, p. 237) in line with Gricean Maxim of Quality. Lackey, in turn, believes that in Stella's case, her assertion will be *misleading* in virtue of the hearer's inference drawn from Quality; she proposes that the supplementary norm of assertion may be proposed to the effect that the speaker is required to assert only what will not be misleading in a context (2007, p. 615).

As we saw, however, in section 1.2.2 of the previous chapter, appealing to the violation of Gricean maxims and explaining Moorean infelicity in terms of conversational implicature conflict is quite hopeless. After all, if Stella were only implying by asserting "Modern-day *Homo sapiens* evolved from *Homo erectus"* that she believes or knows that to be the case, then canceling this implicature (and uttering a Moorean sentence) should *not* mislead anyone, but make perfect sense. Weiner's account similarly fails to steer away from this problem.

Faced with this, one may simply embrace the problematic conclusion and argue that, in fact, such sentences *can* be warrantably asserted, as does Rachel McKinnon[46]:

"Intuitions may differ concerning whether Stella's assertion is too ridiculous to count as warranted. However, one could draw a distinction between warranted assertions (that aren't paradoxical or absurd in the Moorean sense), and warranted assertions that are still paradoxical or absurd in the Moorean sense. One could then say that Stella's assertion is warranted but still strikes us absurd. Now, if we assume that she really is making an assertion, and a warranted assertion at that, then selfless assertions of this kind would be a broad class of warrantedly assertible Moorean sentences. But this is still consistent with the claim that assertible Moorean sentences are rare." (McKinnon 2015, p. 130)

This response is predicated on the idea that the infelicity of Moorean assertion need not be straightforwardly explained by the relevant norm, but may be just the result of them being *rare*. Igor Douven in presenting a case for his own (RCNA) similarly dodges the question of explaining the Moorean linguistic data by saying that "the odd-soundingness of [Moorean assertions] may simply be due to the fact that we never encounter them" (2006, p. 474). The burden of proof now put on the normative account of assertion is not to explain the *unwarrantability* of Moorean assertion, but its infrequency, which may plausibly be accounted for if we assume that *usually* one's assertoric warrant – justification for p – just leads one to believe p and that we *usually* make warranted assertions.

This sort of explanation, however, seems to be far from empirical adequacy concerning mechanisms giving rise to infelicity judgments[47]. Though I conceded in

---

[46] McKinnon, mistakenly in my opinion, ascribes this view also to Lackey and Weiner.

[47] Later on (2009), Douven revoked for this reason his previous explanation of Moore-paradoxicality in favor of a belief-first explanation specified in Bayesian terms. I take this move to imply that Douven now thinks that the impropriety of Moorean assertions is to be explained along the lines

1.5.2 that it is implausible to grant philosophers of language with any "special feeling for inconsistencies" (DeRose 1991, p. 597), it does not mean that all explanations of pragmatic infelicity are created equal. At least since Chomsky's "colorless green ideas" argument, it is widely acknowledged that exposure cannot fully explain our intuitions of grammaticality (Chomsky 1957, p. 15), and similar objections can be deployed against exposure-based accounts of semantic acceptability and pragmatic felicity. A plurality of logically and semantically differing constructions unencountered previously in English can be judged to be typically pragmatically felicitous, though perhaps "clumsy" (such as very long chains of conjunctions), and some less unfamiliar in structure (such as sentences false in most typical contexts, e.g., "It is more than 100°C in here") may be perhaps judged as "odd" if interpreted literally, but still in a qualitatively different way than Moorean ones. There are strong methodological reasons to reject this way of explanation and the authors retorting to it are, in fact, taking on a lot heavier onus of proof if they wish to stand their ground.

One thing worth noting is the more general theoretical tension between an explanation of Moore-paradoxicality in assertion and accounting for Lackey's "selfless assertion" data. Whether or not this tension is resolvable is up to philosophical debate going beyond the scope of this chapter. In my opinion, if one agrees with Lackey's intuitions, a promising strategy would be to argue that in the "selfless" cases, the protagonist expresses an "institutional" belief and asserts as an "occupier of a role" (Sosa 2011, p. 47)[48]. When occupying the epistemic role of a "biology teacher", Stella is then not justified in believing the second conjunct and, hence, cannot properly assert

---

proposed by doxastic norms and that the rational credibility criterion should yield belief in the assertor on pain of irrationality, which makes it consistent with my account.

[48] Interesting experimental data (Turri 2015a) demonstrate that when presented with Lackey's cases, ordinary people tend to ascribe the protagonist of the case both belief and knowledge, thereby supporting a doxastic interpretation of such scenarios that aligns with Sosa's diagnosis.

it. It is clear, however, that taking Lackey's cases at face value is incompatible with a satisfying account of Moore-paradoxicality along the lines set up by the Priority thesis and therefore constitutes a theoretical trade-off.

2.2.2. "Belief without justification" constraints

If avoiding B is impossible, is it possible to avoid the J condition? Among different views on the market, probably the only one suiting the description is the Belief Norm, formulated as follows:

(BNA) *S* may assert *p* only if *S* believes *p*.

BNA may be, in fact, simply treated as a normatively formulated "conventional expression condition" present in our descriptive characterization of assertion; in our context, supporting BNA means simply holding that B is not only necessary, but also a sufficient condition for warranted assertion. Most defenders of BNA, such as Hindricks (2007) and Bach (2008), usually suggest that additional justificatory obligations concerning assertion we seem to have stem from normative constraints put on belief itself – such as the Knowledge Norm of Belief (KNB), which states that one must believe *p* only if one knows *p*. In effect, both Hendricks and Bach suggest that the Knowledge Norm of Assertion (KNA) is just derivable by pairing BNA as a distinctive norm of assertion and the Knowledge Norm of Belief (KNB) as a distinctive norm of belief.

Is this strategy sufficient for a Priority-based explanation or independently plausible? The answer is unfortunately negative. First of all, as mentioned in 1.2, first-person belief reports of the form "I believe that *p*" or "*p*, I believe" usually do not serve as flat-out assertions but as hesitant, *hedged* ones. It seems plausible that the many concrete examples of the following sentence-form can be felicitously asserted:

(BWK) I believe that *p*, but I don't know that *p*.

such as:

(2) I believe that your train will leave in ten minutes, but I don't know that, so better check the timetable.

or:

(3) I don't know, but I believe – Santa comes on Christmas Eve!

On the BNA+KNB view, we get, however, the verdict that such sentences should count as Moorean, and thus be unwarranted. Van Elswyk and Willard-Kyle (forthcoming) observe, that such sentences should be impossible to felicitously assert if both assertion and belief are subject to knowledge norm. This observation may be weakened to hold also if the assertion is only subject to the belief norm, for believing any sentences of this form entails a violation of KNB – and so they cannot be asserted if we agree with Hindricks and Bach. To see that, assume that you know that you believe that *p* and that you do not know *p*. From this it follows that you believe that *p* (since knowledge is factive) and that you believe that you do not know that *p* (since knowledge entails belief). Collection of these beliefs gives rise to a belief in the conjunction of *p* and that you do not know that *p*, which is unknowable. Hence, knowledge of BWK entails that you believe something that cannot be known, which is a violation of KNB – therefore your belief in BWK is improper according to KNB as well. The reason why one cannot assert a Moorean sentence is, according to such a proposal, effectively the same as why one cannot assert BWK. If one finds (as I do) such sentences to be felicitous, then the BNA+KNB view is untenable; but even if one

finds them problematic, it is too strong a claim to say that BWK is Moorean or that it represents the same absurdity[49].

A stronger worry, which entirely defeats the explanatory prospects of BNA, conjoined with KNB or not, is however that it only seemingly derives normative constraints on assertion from similar constraints on belief, and therefore fails to explain why asserting Moorean sentences is defective. For in cases in which an agent *irrationally* believes OMP or CMP and asserts it they may be, according to BNA, only criticized *as believers*, but not *as assertors* – as they properly asserted what they believe, though their belief is irrational (for a similar objection see Montminy 2013, p. 58). Compare this again to Williams' objection to the Priority thesis discussed in section 1.3.: I may be, presumably, perfectly aware that I hold an epistemically irrational belief I just can't shake off and wish to communicate it to you *via* straightforward assertion. BNA cannot give us any reason to regard such assertions as normatively flawed. In effect, a proponent of BNA is left without a good reason to claim that Moorean sentences cannot be properly asserted even if such assertions would express irrational belief.

The same objection will not apply to proponents of stronger doxastic norms, meeting both B and J conditions. If we agree that Moorean belief is essentially irrational, then expressing it through assertion (B) would make one vulnerable to being criticized as an *assertor*, because they are not justified in what they believe and hence not justified in what they assert (J). In the following section, I shall discuss these norms and see how they account for a wider range of Moore-paradoxical data.

---

[49] A similar argument (which replaces "knowledge" with "certainty") may be also found in Hawthorne et al. 2016, pp. 1395-1396.

2.3. Strong doxastic norms

Among the stronger doxastic norms that do meet the preliminary Belief and Justification condition, the following three are most vigorously defended:

(JBNA) *S* may assert *p* only if *S* justifiably believes *p* (Kvanvig 2009).

(KNA) *S* may assert *p* only if *S* knows *p* (Williamson 2000).

(CKNA) *S* may assert *p* only if *S* is epistemically certain of *p* (Stanley 2008).

These norms, as presented, grow in the epistemic demand towards the assertor and (if we understand "epistemic certainty"[50] and "knowledge" in the same way as the original proponents of these norms do) each one is entailed by the next. Given that they all satisfy our preliminary conditions and allow for a unified explanation of Moore's Paradox in belief and assertion along the lines of the Priority thesis, it may seem arbitrary to choose one over the other when it comes to accounting for Moorean data (unless we impose further methodological constraints). Though the account may differ in important details – the proponent of JBNA will explain the impropriety of asserting OMP and CMP in terms of them being irrational to believe, while the proponents of KNA and CKNA will do so by referring to the fact that they cannot be known – they will all agree on that the OMP and CMP are unassertable because of the irrationality or impossibility of the underlying doxastic state[51] and are consistent with the constraint on proper assertion set by the Bridging Principle.

---

[50] Stanley thinks of being epistemically certain that *p* as knowing that *p* and having the highest degree of justification for *p* (2008, p. 35). Therefore, he takes CKNA to entail KNA, not merely require the speaker's maximal subjective certainty.

[51] This picture may become more complicated if one, after Williamson, would like to consider "knowledge" as a *sui generis* mental state, not simply a kind of belief. I will however take this theoretical

2.3.1. Moorean arguments for knowledge and certainty norms

Nevertheless, a closer look at the literature provides us with a variety of significantly different Moore-paradoxical constructions, which do seem at first glance comparably absurd to assert as OMP and CMP, but differ in how the three abovementioned norms account for them. While the proponents of the JBNA-style norms often bring up the infelicity of the justification-version of the Moore-paradoxical constructions as their evidence:

(JOMP) *p*, but I have no justification for believing that *p*.

(JCMP) *p*, but I am justified in believing that ~*p*.

the proponents of the stronger norms report the intuition that the following knowledge- and certainty- versions of Moore-paradoxical constructions[52] are infelicitous:

(EOMP) *p*, but I don't know that *p*.

(COMP) *p*, but I'm not certain that *p*.

Prima facie, the abductive lesson from the examples of JOMP, EOMP, and COMP seems to be that CKNA should be regarded as the most theoretically satisfying. Unlike JBNA, KNA can account for the absurdity of both JOMP and EOMP; it cannot

---

obligation to be independent of the support of KNA; and given that almost all *knowledge-first* theorists (including Williamson) think that knowledge at least *entails* corresponding belief, they would also endorse the Bridging Principle.

[52] Note that, because knowledge is factive (and so is Stanley's epistemic certainty), the *commissive* versions of the below-mentioned examples would be, unlike their doxastic or justification counterparts, simply contradictory.

however, unlike CKNA, account for the absurdity of COMP unless one is committed to the view that knowledge actually requires epistemic certainty, which would make CKNA and KNA equivalent[53]. Stanley utilizes this fact[54] to argue for a methodological superiority of his preferred norm:

"All versions of Moore's paradox (with belief, knowledge, and certainty) would be explained by the invocation of the dual certainty norms for assertion. In contrast, the knowledge account of assertion, according to which assertion is governed by a norm for knowledge, can only explain the belief versions of Moore's paradox, such as [EOMP]" (Stanley 2008, p. 49).

CKNA is not, however, among the most popular views on the philosophical market, and for a good reason – it seems insufferably strong. Simply put, the demand that one should not only know *p*, but do so with indubitable certainty, to perform a seemingly ordinary speech act of asserting *p* is problematic. To see why it is so, we may quickly look at another piece of linguistic data. One of the reasons often given for norms discussed here (especially KNA due to influential argument by Williamson 2000, pp. 252-253) is that they neatly explain the common conversational pattern of challenges to the speaker's assertion in the form of questions: "How do you know that?", "Why do you believe that?" and so on. Such questions crucially differ from the likes of the more aggressive "Do you know that?", "Do you believe that?", "Are you certain of that?" in that in asking them, the speaker presupposes that the assertor *knows*

[53] For such an account see Unger 1975, pp. 259-260.

[54] Stanley also appeals to the fact that asserting "*p*, but it's not certain that *p*" is unwarranted. There are, however, good reasons to not count such conjunctions among Moorean sentences, as discussed in 1.5.1.

or *justifiably believes* the content asserted[55] – and since such presupposition is treated as natural, it might be considered as evidence for one's preferred norm. As Mandelkern and Dorst (2022, p. 14) point out, such questions seem odd if they presuppose that the speaker is epistemically or subjectively certain of the content of their assertion:

(4) There's some beer in the fridge.

(4a.) How can you be certain of that?

(4b.) Why are you so sure of that?

(4c.) Why are you willing to bet your life on that?

If CKNA was true, all of those questions should be proper ways of challenging one's assertion of (4) – but they do not seem so (Mandelkern and Dorst mark them with "#"). Moreover, if it so happens that though I know that the beer is in the fridge (e.g. because I've seen my friend putting six beer-labeled bottles there an hour ago), I am not epistemically certain of that (e.g., because I was not present in the kitchen the whole time or because I did not inspect that the bottles did not contain milk upon my friends' arrival) it seems wrong to call me out and expect me to retract my assertion – at least if one is not David Hume or Pyrrho.

While Mandelkern and Dorst take such conversational patterns to be sufficient to establish that CKNA is false, they also maintain that COMP constructions are typically infelicitous. But is that right? Let us assume the context associated with (4) in the last paragraph. I will concede that the following out-of-the-blue assertion seems defective:

---

[55] Assuming, as it is customary, that "How *p*?" and "Why *p*?" questions presuppose the truth of *p* (Bromberger 1966).

(5) There's some beer in the fridge, but I'm not certain of that.

But now consider the following versions:

(6) There's some beer in the fridge, but I'm not certain of that, *since I was not standing in the kitchen all the time; however, I see no good reason why it should disappear.*

(7) There's some beer in the fridge, but I'm not certain of that, *since I did not check what's in the beer bottles; however, I see no good reason why it should be something else than beer.*

One may also produce a similar effect by putting an emphatic stress[56] on the adjective "certain":

(8) There's some beer in the fridge, but I'm not *certain* of that, since…

or (as Williamson 2000, p. 254 humorously points out) by descriptively fixing the certainty standards incredibly high:

(9) There's some beer in the fridge, but I am not certain of that *by Descartes' standards*.

Are all of these assertions as bad as (5)? I take it to be intuitive that they are not[57]. What's the crucial difference between them? The reason here seems to me largely pragmatic – for unlike in cases discussed in section 2.2, the italicized provisos play a role that can be easily likened to implicature cancellations of the second conjunct. Consider the fact that usually, when we say that we are *not* certain of something ("I

[56] A similar argument from emphatic stress is employed by Hintikka 1962, p. 100.

[57] For this claim not to remain unfounded: these data are also contested by van Elswyk and Benton (2023, pp. 35-36), who cite McCready's (2015) classification of such constructions as "shield hedges" and provide real-life examples of similar assertions from speech corpora.

am not certain of *p*"), we do not mean to take into account the doubts of a radical skeptic but rather communicate that we are in an imperfect epistemic position with respect to *p*. In such cases, the added proviso, italicized in the examples, seems to secure the epistemic position of the speaker required to assert that *p* by generally following the pattern of making it clear that one does not meet the "high" standard (*since…* part), while simultaneously meeting the sufficient one (*however…*). If one may make sense of such provisos and regard them as properly defending the original assertion, then I can see no reason why the likes of (5) should be regarded as absurd for more than pragmatic reasons in most ordinary cases.

2.3.2. Defensible and indefensible epistemic Moorean assertions

Is this observation, that is that one may coherently defend their assertion of COMP, limited to *certainty* ascriptions, or can be pushed further? If my explanation of why such defenses work is right, then it *cannot* be extended to JOMP: for I still need to provide some reasons for *p* being true to defend my warrant to assert that *p*. But what about EOMP? Unlike in the certainty wording, the abovementioned example may seem felicitous in the "knowledge wording" only in the mouth of someone who holds a certainty-high standard for knowledge. However, one may find plausible cases in which one may permissibly defend one's assertion of EOMP when the subject matter is typically treated as epistemically inaccessible or at least sufficiently hard to access. For example, Hintikka (1962, p. 100) cites the following assertion of EOMP:

(10) God is almighty, although I don't know that He is.

as not absurd, or at least comparably less so than its version containing an empirically verifiable factual claim:

(11) Phosphorus melts at 44°C, but I don't know that it does.

For a less controversial case than religious assertions (which sometimes had been described as cases of *expressive* rather than assertoric speech acts), one might consider a more familiar discourse regarding future contingents[58]. Consider the following (again adopted from Hintikka 1962, p. 100):

(12) Labour will win the next elections and form a majoritarian government; of course, I don't *know* that, *since I have no insider information, but all the data and trends in polling suggest that.*[59]

Weiner (2005, p. 239) also presents a similar case of responding to a "How do you know that?" challenge to an unqualified assertion regarding future events which intuitively seems admissible:

(13) I don't *know* if the French will attack at nightfall – we haven't intercepted their orders – but my prediction is that they will.

A proponent of KNA or CKNA may disagree with classifying such examples as felicitous[60] or deny that predictions or expressions of faith should be held to similar standards as "ordinary" assertions. The second option seems to go well with the observation that such cases of felicitous Moorean assertions are d*iscourse-specific;* for the admissibility of Hintikka's or Weiner's examples and the defensive strategy offered

[58] Another similar and interesting case concerns the discourse about plans and practically rational future actions. See Maitra and Weatherson 2010.

[59] This sentence was tokened in this file in 2023.

[60] For example, Benton writes of examples given by Weiner that their explicit formulations in K-OMP form still "sound quite bad" (2012, p. 103). I will not argue against this view; let me only say that this intuition seems to be an intuition of the minority, and more authors take these constructions to be admissible (apart from works cited earlier see e.g. Williams 2023).

above does not provide us with a general recipe to defend *any* EOMP assertion. Not to engage in controversial examples, one may think here of mathematical assertions. Given a plausible picture of mathematical discourse, there is simply no way to defend:

(14) Pythagorean Theorem is true, but I don't know that.

for to do so, one needs to cast doubt on whether the justification one has for believing the theorem is sufficient for knowledge. If the only admissible justification for mathematical assertion (in contrast to hypothesis or conjecture) is proof, and we also take having proof to be sufficient for mathematical knowledge (see Williamson 2000, pp. 263-266), then any defense one may produce for the second conjunct at the same time defeats one's warrant for asserting the first. Similarly, one may think of examples concerning avowals of direct sense-perception or belief possession, where even COMP constructions seem to have no appropriate defense:

(15) I see A, but I'm not certain that I see A.

(16) I believe that p, but I'm not certain that I believe that p.

This asymmetry of the possibility of defense suggests that we should seek a more flexible account of what goes off with Moorean assertions and their close cousins such as predictions. For even if knowledge or certainty is *normally* required for assertion, the fact that predictions such as:

(17) Labour will win the next elections and form a majoritarian government, but I don't believe that.

(18) Labour will win the next elections and form a majoritarian government, but I am not justified in believing that.

remain indefensible in a way in which their knowledge or certainty counterparts are not, requires an appeal to a weaker norm. Though the proponent of CKNA or KNA may hope to delineate assertions *proper* as speech acts subject to the knowledge or

certainty norm and classify the remaining cases such as predictions, religious constatives, and so on, as different speech acts subject to weaker norms, this strategy seems unattractive. For one, it suffers from an objection of being *ad hoc* – such a theory of what assertions are remains closed to any possible falsification attempt and can be hardly regarded as explanatory. Secondly, the term "assertion" is descriptively better thought of as force- rather than content-specific. The difference between, e.g., asserting *p* and hypothesizing *p* is usually drawn along the lines of what level of conviction is needed from an agent or what type of belief about *p* they need to have (whether they need to believe *p* or just believe that it would be good to consider consequences of *p*) to perform the relevant speech act, not what type of content *p* has.

A more promising response would be to point out that *predictions* are somewhat special cases of assertion that may be subject to *discourse-specific* norms. As previously noted, the same strategy could be applied to expressions of faith. It can also be extended by an observation that other constative[61] content-individuated speech acts, such as moral assertions were given analyses that put *stronger* conditions on the speaker's epistemic state (see, e.g., Kelp 2020) without denying them the status of assertions[62]. As Hintikka observes in such examples[63]:

"This suggests that the absurdity of [EOMP] is not due simply and solely to the impossibility of knowing the truth of what [EOMP] expresses. Its absurdity is partly due to absence of any indication of *some special circumstances which would relieve one from the normal expectation that one can know what one is saying*, as well as to the absence of any disclaimer that

---

[61] In Bach and Harnish's terminology (1979, p. 41); Searle uses the term (perhaps more apt in our context) "assertives" (1976), while Turri (2012) proposes a term "alethic speech acts".

[62] See also Gordon's (2023) recent proposal concerning assertions in political discourse.

[63] Cf. Williamson (2000, p. 254) on C-OMP constructions.

would indicate that the speaker is alive to the fact and ready to admit that he cannot know what he is saying. (…)" (Hintikka 1962, p. 99, italics mine).

A natural thing is then to propose that assertion itself is not guided by any particular norm but rather a set of different norms applicable to different topic-individuated discourses, such as future-directed, moral, aesthetic, scientific, or philosophical discourse, with knowledge being a "default" expectation towards speakers which could be either heightened or lowered depending on the specific rules of the respective discourse[64]. In the next section, I shall sketch such an account and demonstrate how it allows us to predict correct felicity verdicts for a variety of different, Moore-paradoxical statements.

2.4. Flexible Assertion Schema

How should we approach the task of designing the minimal flexible account of assertion which allows us to dispose of different Moore-paradoxical constructions while retaining the idea that some constructions with an identical structure are, in fact,

---

[64] The plausibility of such an account is, in my view, quite independent of whether one prefers to distinguish descriptive and normative enterprises as I do or embrace the Williamsonian constitutive-rule view of such norms. If you like a bit worn-out analogy between linguistic practice and chess, you may think of how *castling* gets defined across different variants of the game, like 960 Chess (Fischer chess) or Capablanca chess, or how it *was* defined earlier in the development of the game before adopting modern conventions. Though in such variants the king and the rook perform different moves to castle than in the traditional version of the game, the similarity and the aim of the move, as well as the overall strategic pattern of the game, are sufficiently similar to its classical understanding to say: "Castling in 960 Chess works slightly differently than in the standard chess" rather than "960 Chess have a different move that's also called 'castling'". It seems to me inappropriate to say without some heavy qualifications that these different norms of castling are norms of *different moves*, or that one of such norms (say, of castling in traditional chess) is especially privileged rather than simply most salient and frequently employed.

admissible? To stay complicit with the Bridging Principle, the "lower bound" of such an account needs to require that the assertor of *p* justifiably believes that *p* (where the justification of one's belief that *p* minimally meets the threshold for rational belief that *p*). This goes in line with the observation made in the previous section that while some COMP or EOMP constructions are defensible, that is not the case with simple CMP, OMP or JOMP constructions – while "Labour will win, but I don't know that" may be defended, "Labour will win, but I am not justified in believing that" or "Labour will win, but I believe they will not" cannot, given the restrictions of predictive discourse.

One such flexible account[65] which nevertheless remains true to the core idea that assertion is a distinctively epistemic speech act is a context-sensitive proposal of Goldberg (2015). He defends the view of assertion "which regards the core rule for assertion as requiring one to have the relevant epistemic authority, but where what counts as the relevant authority will vary according to what is mutually manifest (...) in context." (2015, p. ix). According to Goldberg, this context-sensitivity is borne by the conflict between the existence of contexts in which the widespread presence of peer disagreement may act as a defeater for one's knowledge that *p* and, on the other hand, the continuous need to engage in assertoric practice with respect to *p*. In turn, without abandoning the central idea that assertion is a primarily epistemic speech act, he proposes the following:

"When it comes to a particular assertion that p, the relevant warranting authority regarding p depends in part on *what it would be reasonable for all parties to believe* is mutually believed among them (regarding such things as the participants' interests and informational

[65] Other context-sensitive accounts are McKinnon (2013, 2015) and Gerken (2012). As both of these proposals are, however, cases of non-doxastic norms, they are susceptible to critique developed in section 2.2.1.

needs, and the prospects for high-quality information in the domain in question)." (Goldberg 2015, p. 266)

Though Goldberg's case there is mainly centered on *philosophical* disagreement, we might very well intuitively extend this to discourses – individuated as sets of speaker's expectations and epistemic prospects (as suggested by Goldberg) concerning some topic or "domain" – discussed in the previous section. Unlike him, I am not, however, convinced that the presence of such disagreement might defeat one's rational belief, though surely it undermines the possibility of *knowledge* – as indicated in section 2.2.2. (and as shall be later extensively defended in chapter 5.), we have independent reasons to think that one may rationally take oneself not to know that *p* while maintaining rational belief in *p*. My idea to operationalize Goldberg's idea for our present purposes is to instead utilize the descriptive characterization of assertion as a conventional expression[66] of belief (and a warranted, proper assertion – its actual expression) and argue that the normative expectations of the hearers towards its justification are fixed by relevant features of a given discursive practice. We might capture this by postulating the following schema as yielding different specific norms of assertion relative to different discourses:

(Assertion Schema) S may assert *p* in discourse *D* only if S's belief that *p* expressed by such assertion meets justification standards relevant for *D*.

[66] Another theorist who defends the *expression* condition in the normative setting is Turri 2011, who notices that in its absence one's "lucky random assertions" (p. 40) may be warranted if one non-occurently knows or rationally believes *p* when they assert.

In the next two subsections I will demonstrate how this schema may be used to generate specific norms guiding assertoric practice within a certain discourse.

3.4.1. Applications of the Flexible Schema: varying justification strength

The most important question, of course, is: what sort of "discourse-relevant standards" are we speaking of here? I think it is useful to distinguish and discuss two potential restrictions: that of *strength* and the *kind* of justification required in the relevant discourse. Let me start by focusing on strength restrictions. Notice, following Hintikka's observation mentioned in the previous section, that in most ordinary circumstances the relevant constraint on the strength of one's justification is the strength allowing the belief in question to qualify as knowledge (hence validating the basic intuitions underlying the support for KNA), while in other cases the strength of justification required is either lower (e.g., for predictions) or higher (e.g., for avowals). This way, in line with the observations made in the previous sections, we may easily account both for the intuitive infelicity of EOMP when uttered in the discourse concerning intersubjectively verifiable matters (such as (11)) with its admissibility in the discourse concerning facts not verifiable this way, such as discourse involving future contingents (as per Weiner's and Hintikka's examples). To capture this idea, we might think of the following precisification of Assertion Schema for sentences with future-directed content:

(FNA) *S* may assert *p* (where *p* is a future contingent sentence) in the future-directed discourse only if *S*'s justification for the belief that *p* expressed by such assertion is based on a reliable predictive mechanism.

Let us look how such a norm could explain the felicity intuition behind (12) and (13). According to FNA, one may, without claiming to *know* p, assert *p* if they are justified in believing that Labour will win the next election and form a majoritarian

government based on reliable polls and assessment of Conservative's poor standing, but that such belief does not meet requirements for being knowledge (such as being factive, if we allow for the future to be open, being causally related to the predicted event or otherwise). I wish not to enter here the debate on what exact type of justification is proper for believing future contingent statements – the exact wording of FNA would undoubtedly at least depend on whether internalist or externalist views on justification are true. This claim should be read only as a rough proposal of how such discourse-specific norms would look like, in this case from an epistemic externalist perspective, while its accurate wording should, of course, rely on careful epistemological considerations regarding our epistemic toolkit for making accurate predictions.

Analogously, stronger conditions related to the content of *S*'s assertion may be imposed, e.g., for avowals:

(AvNA) *S* may assert that *p* (where *p* predicates (not) being in *M* of *S*) in the first-person discourse only if *S*'s justification for the belief that *p* expressed by such assertion makes *S* epistemically certain of *p*.

Unlike predictions, avowals tend to be held to much higher standards of subjective certainty than regular assertions, and much more often than they are taken at face value rather than challenged. Until Chapter 5., I shall not go into the dispute of what type of justification is the source of regularly noticed peculiar safety of avowals and second-order beliefs, however, AvNA is convergent with the standard observation that avowals enjoy, in general, a privileged position among other assertions in terms of the authority of the speaker[67]. AvNA also predicts, in line with

[67] One may also observe here that the regular "How do you know that?" or "Why do you believe that?" challenges seem to be out of place when it comes to challenging avowals; this inappropriateness,

the intuitions of many philosophers, that mental state self-reports based on less-than-certain evidence seem peculiarly out-of-place, such as one's self-ascription of pain based purely on the observation of one's pain behavior.

Accepting AvNA easily vindicates the intuition that the likes of (15) and (16) are, unlike those with future-directed content, indefensible. If we allow for different discourses to determine their justification-strength criteria and allow for multiple realizations of such norm of assertion, content-specific differences in felicity judgments concerning Moore-paradoxical claims are easily validated.

3.4.1. Applications of the Flexible Schema: varying justification kind

Such flexibility also allows us to posit that some discourses specify not only the strength but also the *kind* of justification admissible. One of such constraints, widely discussed in the literature on aesthetic and taste judgments, is *direct acquaintance* with the object of judgment. Consider, for example, the following sentences:

(19) *The Starry Night* is beautiful, but I have never seen it.

(20) *The Starry Night* is beautiful, but I dislike it.

Asserting (19) or (20) seems to be infelicitous in a familiar Moore-paradoxical way in which (21) seems not:

(21) Phosphorus melts at 44°C, but I have never seen it melting.

---

however, seems to stem much more from the assumed security of the truth and justification of avowals rather than the fact that we don't ascribe knowledge of one's mental state to the assertor.

While this contrast is usually brought up to highlight the Kantian point that first-hand acquaintance is necessary for the formulation of appropriate aesthetic judgment, we may think of it also as a property of aesthetic discourse and what can count as a warranted assertion in such disputes. Though merely believing that *The Starry Night* on the basis of someone else's testimony (say, that of a reputable art critic) plausibly would not qualify as irrational, expressing it through assertion gives rise to widely observed and corroborated "acquaintance inference" (Ninan 2014). John Collins (2021, p. 978) proposes that the following, broadly Kantian norm (which he dubs NAA) may guide our aesthetic ascriptives[68]:

(NAA) *S* may assert that *x is beautiful* only if *S* takes pleasure in the experience of *x* that is not based upon idiosyncratic features or etiology of the experience, and so the judgment is to be commended universally as based upon *S*'s experience of *x* independent of all other factors peculiar to *S*.

If we try to reformulate this norm to get the instance of the flexible schema, we should get the following norm specific to the aesthetic discourse:

(ANA) *S* may assert *p* (where *p* predicates *beauty* of *x*) in aesthetic discourse only if *S*'s belief that *p* expressed by such assertion is based on

[68] In Tarnowski 2024 I also sketch a proposal based on Collins' norm for aesthetic assertion for assertions concerning taste, which is another discourse in which direct acquaintance is frequently presupposed.

> *S*'s pleasure in the experience of *x* that is not based upon idiosyncratic features or etiology of the experience, and so the judgment is to be commended universally as based upon *S*'s experience of *x* independent of all other factors peculiar to *S*.

Given ANA, the first conjunct in both examples, if warranted, expresses a belief that is based on the pleasurable experience of *The Starry Night*, while the second – knowledge that one has not seen *The Starry Night* or that one dislikes it, which is jointly incompatible. We have then good grounds for assimilating such peculiar assertions to a wider class of Moore-paradoxical ones, while still maintaining that the nature of the paradox is, in general, belief-based in line with the Priority Thesis.

The cited examples (FNA, AvNA, ANA) are of course not exhaustive, as one may easily find a variety of other cases where specific, discursively demanded justification constraints may apply[69]. I am also not explicitly defending the wording nor specific conditions imposed in any of these concretizations of the Assertion Schema. The point of this section is that allowing such flexibility should be welcome, given the different intuitions we have when it comes to the defensibility of different Moore-paradoxical constructions; given the fact that Moore's Paradox obtains not only in assertion but also in belief, the borders of this flexibility should be, however, fixed with JBNA as the lower bound.

One might object that the schema's "flexibility" designed to filter out relevant non-absurd Moore-paradoxical assertions is somewhat *ad hoc*. In the above considerations, the apparent felicity or infelicity of different Moore-paradoxical

[69] As noted above, religious and moral discourse may be obvious cases. See also Gaszczyk (2022) for a comprehensive review of a variety of norms formulated for claims with different content and other constative speech acts.

constructions seems to play a double role as both the schema's testing device and its *explanandum*. While we may, in response, frame the argumentation in this section as essentially abductive – that is, proposing the best (or at least a better) explanation of variance in infelicity judgments concerning Moore-like assertions – this still may leave us wondering whether there is any independent reason to impose such discourse-sensitive distinctions? And is there any principled way of doing so without appealing to judgments of felicity and infelicity?

I think that an initial pull against such discourse-specificity may come from the intuition that an assertion as such needs to have a distinctive aim individuating it from other types of speech acts uniform across different contexts. One may go on to suggest (e.g., after DeRose 2002) that the relevant contextual differences should be accounted for by making our *knowledge* ascriptions contextually sensitive (and, e.g., maintaining uniformly KNA across the board), rather than allowing the norm of assertion to be flexible; a theorist sympathetic to a weaker norm may plausibly make a similar move by making *justified* contextually sensitive term.

While I am not entirely unsympathetic to such theoretical moves, I think the contextual flexibility of norms of *assertion* seems to better fit the idea that such norms of assertion's normative propriety are (at least partly) established in sensitivity to the linguistic practice they play the role in. Consider the different discourses we covered in the last section: future-directed, introspective, and aesthetic, with discourses about empirically directly verifiable facts. What differentiates them is (among other things) the epistemic position of assertors and the audience alike concerning the discourse's topic. The epistemic environment – that is, the evidence *in principle* possible to obtain by the participants of the discourse – is radically different both quantitatively (the number of different pieces of evidence to be gathered and its decisiveness) and qualitatively (the admissible sources of such evidence). Unlike in the standard case of discourse concerning empirically verifiable facts, the evidential position of discourse participants may also be, by default, asymmetrical, as is with the case of introspective

discourse. In cases of aesthetic and future-directed discourses, the irresolvable disagreement among epistemic peers, serving as a main motivation for Goldberg's proposal, is also widespread, which plays an important role in how speakers ought to evaluate their respective justification (see: Goldberg 2015, pp. 231-244). While this all may be, in principle, spelled out solely in terms of discourse-sensitive standards for knowledge or justification instead of assertion, it is also pretty reasonable to assume that the differences in the "harshness" of the epistemic environment may give rise to different discursive practices which, in turn, constitute what counts as a warranted assertion in the given inquiry. Moreover, if all participants recognize that in some specifically harsh or limited inquiry one usually cannot claim to *know* or *be certain* that *p*, it would make sense to reorganize the rules of the discourse accordingly, although with the same aim (making a progress in a collective inquiry) in mind, rather than pretend that in such a case knowing that *p* is possible despite one's relatively poor epistemic standing. If admitting different types of epistemic warrant for an assertion to be proper across epistemically diverse discourses may be rational, then the flexible account sketched above seems plausible independently of the considerations concerning Moore-paradoxicality and hence may be regarded as explanatory.

The idea defended here is, of course, not a complete theory of warranted assertion, but rather it's embryonic form. A fully fleshed out account along these lines should offer a precise taxonomy of different discourses and specify the principles of tying them up with epistemic expectations of speakers. As this dissertation is concerned primarily with an explanation of Moore-paradoxicality, I limit myself here to saying that discourse-specific variance in felicity judgments concerning Moorean sentences is at least an indication that such an account is needed. I shall return to this issue briefly in Chapter 6.; nevertheless, I believe that the above considerations make a convincing case for the claim that to explain infelicity of Moorean assertions, one needs a fine-grained account of normative constraints on assertion that is sensitive to shifting evidential standards.

2.5. Conclusion

In this chapter, I analyzed which of the various accounts proposed in the literature does best when it comes to explaining the infelicity of Moorean assertions. Following the methodological constraint specified in the previous chapter, I argued that a satisfying account should make a proposed analysis continuous with an explanation of the irrationality of Moorean belief along the lines of the Priority thesis and justify the reformulated Bridging Principle, linking assertion with belief. In 2.1. I distinguished normative and descriptive accounts of assertion. In section 2.2., I defended the two constraints flowing from the Bridging Principle: that for an assertion to be warranted it needs to express belief (B) and be appropriately justified (J). By commenting on the shortcomings of norms of assertion denying either B (Weiner's TNA, Lackey's RTBNA) or J (Bach's and Hendrick's BNA) conditions, I demonstrated why these conditions are plausible and should be expected of any normative account of assertion which takes upon the task of explaining judgments of Moorean infelicity.

After setting up these preconditions, I proceeded to analyze the most prominent strong doxastic norms in section 2.3: the Justified Belief (JBNA, Kvanvig 2009), Knowledge (KNA, Williamson 2000), and Epistemic Certainty (CKNA, Stanley 2008) norms. I argued that upholding any particular norm of the three leads to trouble in accounting for the fact that stronger Moorean constructions involving "know" (EOMP) or "being certain" (COMP) seem to differ in their felicity and defensibility across different contexts. In section 2.4. I proposed a modest modification of a strong doxastic norm in the form of a context-sensitive schema, which subjects the expected strength and nature of the speaker's justification for their belief in the assertion's content to norms demanded by the discourse within which the assertion is made. Such a "flexible" account was shown to give rise to different, discourse-specific norms of assertion

(future-oriented, introspective, aesthetic) predicting the right verdicts of defensibility of Moore-paradoxical constructions.

Chapter 3. Non-Assertoric Moorean Speech Acts

The explanation of the infelicity of Moorean assertions offered in the previous chapter was based on the idea that the assertoric practice is subject to specific epistemic constraints. I hypothesized that these constraints essentially vary over different discourses, which guide expectations towards the strength and kind of justification the speaker has for what they assert. However, one might rightly wonder whether this idea captures everything about Moore-paradoxicality in speech. Isn't focusing entirely on the *assertoric* practice a case of narrow-mindedness, stemming from the "fetishization" of the speech act of assertion common in the analytic philosophy of language (McGlynn 2014, p. 82)?

Although the issue is often side-lined in the discussion on Moore's Paradox, it has been observed that phenomena eliciting intuitions of "Moorean" infelicity are not limited to assertions or other constatives. Consider the following examples discussed in the literature:

(Moorean Promise; MP) I promise I'll marry you, but I won't.

(Woods 2018)

(Moorean Command; MC) I order you to turn in your final paper by the end of the exam period, but you might not turn it in by then.

(Mandelkern 2021)

(Moorean Request; MR) Close that door! I don't want you to do it, however.

(Black 1952)

(Moorean Question; MQ) How old are you? – but I don't want you to tell me.

(Shoemaker 1988)

(Moorean Apology; MA) I apologize, but I don't feel contrition for what I did.

(Allston 2000)

These constructions present a puzzle for a Priority-based analysis of Moore-paradoxicality in terms of expressing inconsistent or otherwise peculiar beliefs and the standard treatment of speech acts, which does not classify promises, commands, or requests as expressing beliefs. If we commit to the view that *all there is to Moore-paradoxicality in speech* is adequately captured by a Priority-based explanation, (partly[70]) non-assertoric speech acts exhibiting this feature should strike us as peculiar. As Jack Woods explains:

"The priority thesis (...) is insufficient to give a general explanation of Moore-paradoxical phenomena (...), as [MP] has the same paradoxical character as [CMP], even though the latter, but not the former, involves expression of the utterer believing thus and so. Moore-paradoxical utterances can occur in non-assertoric constructions and, moreover, can occur in the context of speech acts which do not express mental states at all" (Woods 2018, pp. 323-324).

Contrary to Woods, I don't take this observation to be a decisive counterexample, but rather an inspiring challenge for the Priority-based explanation. A natural strategy for a Priority theorist is to plausibly explain why, despite appearance, the use of such constructions may signify irrationality at the level of belief which is normatively unexpected of speakers performing commands, promises, or apologies. This would

[70] Note that all of the mentioned examples nevertheless require pairing non-assertoric speech acts with *assertions.*

make the analysis of non-assertoric Moorean speech acts continuous with the approach to Moorean assertion proposed in Chapter 2.

The problem is that, as of yet, we do not seem to have any similar, normative treatment of non-assertoric speech acts at hand (Harris et al. 2018, pp. 12-13). This chapter undertakes this task, building upon the framework of Stalnakerian dynamic pragmatics. In section 3.1, I will offer a taxonomy of different examples proposed in the literature and discuss the theoretical desiderata specific to the Priority-based explanation of non-assertoric Moorean infelicity. Section 3.2 will develop a central hypothesis of this chapter, and demonstrate how it allows accounting for the infelicity data concerning Moorean promises and commands; section 3.3 will discuss extending this view to cover other directives and commissives as well as other types of speech acts. In section 3.4 I will finally discuss two controversial pieces of data: infelicity arising in conversational pretense and selfless speech acts.

3.1. Preliminaries

Before getting into the Priority-based analysis of these puzzling examples, introducing some background theoretical framework and making preliminary distinctions is needed. In this section, I will first (in 3.1.1.) introduce an approximate classification of speech act types based on Bach and Harnish's (1979) taxonomic principles which distinguishes them by the type of mental state they conventionally express (in the sense introduced in the previous chapter). This will allow me to provide a fine-grained taxonomy of Moorean speech acts. Subsection 3.1.2. will lay out some constraints this categorization puts on our analysis. In 3.1.3. I will discuss the prospects of providing an account continuous with the account of the Moore-paradoxical assertion defended in Chapter 2. and introduce the guiding principles of Stalnaker's dynamic pragmatics, which will become useful in formulating my account.

3.1.1. Taxonomy of Moorean speech acts

The first thing we should observe is that, despite eliciting a similar feeling of "Moorean absurdity" or infelicity, the five examples provided in the introduction are quite diverse. For starters, they seem to cover pretty much all possible types of speech acts: Moore-paradoxicality seems not only to be confined to assertions but also conjunctions of promises, requests, questions, commands, apologies, and so on with a subsequent assertion. The required analysis – if we agree that it should be uniform – needs then to take into account what is common in all these cases and abstract from peculiarities connected with this or the other specific type of speech act.

Following John Searle's (1970, 1976) and Bach and Harnish's (1979) taxonomies of speech acts[71], we may classify them into five broad categories: constatives (e.g., assertions, hypotheses, guarantees), directives (e.g., commands, requests, questions[72]) commissives (e.g., promises, threats), expressives (e.g., apologies, good wishes, curses), and declarations[73] (e.g., baptisms, marriages, decrees). Following more closely the latter treatment, we may say that the crucial difference between these categories lies in the type of mental state they are conventionally expressing: belief for

---

[71] In general, I assume that one may faithfully translate what is written in this chapter to one's preferred terminology or taxonomy of speech acts, provided that it maintains the link between different kinds of speech acts and types of attitudes conventionally expressed by their performance.

[72] Following Searle (1976, p. 11, f. 2) and Bach and Harnish (1979, p. 48) I tentatively include questions in the directive category, as requests for information, expressing the desire to know or to obtain information from the hearer. Though it's perhaps crude, it will do for the present purposes of exposition. See van Elswyk (2023) for a general contemporary account of questions in this direction.

[73] I borrow the first three names from Bach and Harnish and the second two from Searle, based on their entrenchment in the relevant literature (e.g., speaking of "constatives" is far more popular than of "representatives").

constatives, desire for directives, intention for commissives, reactive attitudes[74] for expressives and none for declarations. Further distinctions within these classes (e.g. between a promise and a threat within commissives) can be approximated by taking into account the specific relation between the attitude and its content (e.g., a promise is taken to express an intention to φ, while a threat an intention to φ conditional on the hearer not ψ-ing).

This classification is plausibly insufficient to deliver a full and precise taxonomy, as it does not account, e.g., for specific social placement of the speaker and hearer (which, e.g., allows us to fine-grainely distinguish commands from requests) or strictly conventional differences between speech acts. Nevertheless, as Searle writes: "[i]f one tries to do a classification of illocutionary acts based entirely on differently expressed psychological states (...) one can get quite a long way" (1976, p. 4), and this is as far as we need to go for the purposes of this chapter; focusing on further subtleties would only obscure the main point. The following table presents the basic principles of such rough classification:

[74] Bach and Harnish speak here of "certain feelings" (1979, p. 51). I borrow the term from Strawson (1962) to speak broadly of feelings such as anger, contrition, gratitude, love, etc., to underline the fact that they are non-propositional and that they are usually directed towards some person or event (for a similar reason, Bach and Harnish label expressives "acknowledgments").

Speech act category	**Attitude kind**	**Speech act type**	**Expressed attitude**
Constatives	Belief	Assertion (that *p*)	Belief that *p*
		Conjecture (that *p*)	Belief that it is worth considering the consequences of *p*
		...	...
Directives	Desire	Command (that *H* φ′s)	Desire that *H* φ′s
		Question (to *H* whether *p*)	Desire that *H* tells *S* whether *p*
		...	....
Commissives	Intention	Promise (to φ)	Intention to φ
		Threat (to *H* that *S will* φ unless *H* ψ′s)	Intention to φ conditional on *H* not ψ-ing
		....	...
Expressives	Reactive attitude	Apology (for φ-ing)	Contrition (that *S* φ-d)
		Thanks (for φ-ing)	Gratitude (that *H* φ-d)
		....	...
Declarations		Wedding	
		Baptizing	
		...	

Table 1. Speech act classification by the type of expressed attitudes

The five opening examples seem to cover pretty much all of these categories but declarations. Yet, it is easy to imagine possible cases of Moorean constructions for declarations, such as:

(Moorean Marriage; MM) I now pronounce you man and wife, but you're not married.

If we agree that this example is genuinely problematic and not, e.g. simply contradictory or not a marriage at all, Moore-paradoxicality can be then exhibited by pairs (conjunctions or subsequent utterances) of all different types of speech acts with certain assertions.

The second observation we should make is that there is a plurality of different types of assertions that, conjoined with a specific speech act, give rise to Moorean infelicity. For example, in the Moorean Request case, the request for the hearer to φ is then followed by a *disavowal of a corresponding desire for the hearer to* φ; but in the Moorean Promise, the second conjunct states something directly about the success of previously made promise and nothing about one's intentions. Building on this, we may roughly distinguish at least two types, which I shall label "Sincerity" and "Satisfaction" formulations. Sincerity formulations are formed of pairs of a respective speech act and the assertion concerning its sincerity condition, i.e. the mental state *M* conventionally expressed by this very speech act. OMP, but also MR, MQ, MS and MA will count as belonging to this group, as in all of these cases we encounter such disavowals. Satisfaction conditions, on the other hand, are formed of pairs of speech acts with the assertion concerning the *satisfaction* condition of *M*, i.e. the truth of the expressed belief, the fulfillment of the expressed desire, or the execution of the expressed intention. MP and MC will belong to this category, as the assertoric parts forming them concern the satisfaction conditions of expressed intention and desire.

Tentatively, we might also include MM in this category – although declarations do not conventionally express *any* mental states (for this reason they do not form Sincerity formulations), they do have certain satisfaction conditions of their own, i.e. intended institutional or social effects their performance has on the world. The same will not apply to expressives, as they do express attitudes (contrition, happiness, etc.) that do not have satisfaction conditions.

Thirdly and lastly, we should note that both Sincerity and Satisfaction formulations will exhibit the duality of the *omissive* and the *commissive* form, noticed already in the classic OMP-CMP case. When it comes to Sincerity formulations, the difference between the forms can be spelled out thusly: the omissive forms will pair the speech act with the *disavowal* of the respective expressed mental state, while the commissive forms with the *avowal* of a contrary mental state, e.g.:

(1) Close the window, but I want you to leave it open.
(2) I promise to marry you, but I intend to not marry you.
(3) I apologize for stamping over your foot, but I feel glad that I did.

In the case of Satisfaction formulations, we may observe a similar duality when it comes to whether the assertion concerning the satisfaction condition *s* of *M* explicitly *denies* that *s* does not or will not obtain or merely *leaves it open*, e.g. by prefixing the denial with an epistemic modal verb ("it may be that ~*s*"). As Matthew Mandelkern observes, when we pair orders with such weak assertions, as in MC or:

(4) You should turn in your final paper by the end of the exam period, but you might not turn it in by then.

we still observe a familiar feeling of Moorean infelicity[75]. A similar observation naturally extends to other speech acts, such as commissives or declarations, as witnessed in the following two cases:

(5) I promise I will marry you, but I might not marry you.
(6) I pronounce you man and wife, but you may not be married.

To maintain terminological uniformity, I allow myself to put the more restrictive constructions containing denials into the class of commissive, and the "open" constructions into the class of omissive forms of Moorean speech acts.

The summary of these classifications – specifying the type of speech act in the first conjunct, the form of the second one, and the formulation of the pair – is presented in the following table. To maintain uniformity and abstract from the unnecessary complications with interpretation which would now obscure the intended aim, I use explicit performative wording ("I α that...", "I α you to...") for the first[76] conjunct of the speech act:

[75] If the first conjunct of the sentence is interpreted as an order, and uttered in the imperative mood.

[76] Of course, I refer to the "first" and "second" conjunct for convenience; the order of the conjuncts does not matter (conjunctions "I do not intend to marry you, but I promise to" or "You will not close the window, but I order you to" are not less infelicitous than their reversed forms).

Category of a speech act	**Sincerity formulation** *(paradigmatic example)*	**Satisfaction formulation** *(paradigmatic example)*
Constative (omissive and commissive)	I assert that *p*, but I don't believe that *p.*	I assert that *p*, but it might be that *~p.*
	I assert that *p*, but I believe that *~p.*	I assert that *p*, but *~p*[77].
Directive	I command you to φ, but I don't want you to φ.	I command you to φ, but you may not φ.
	I command you to φ, but I want you not to φ.	I command you to φ, but you will not φ.
Commissive	I promise to φ, but I don't intend to φ.	I promise to φ, but I may not φ.
	I promise to φ, but I intend not to φ.	I promise to φ, but I will not φ.
Expressive	I apologize for φ-ing, but I don't feel contrition for φ-ing.	
	I apologize for φ-ing, but I am glad that I φ-d.	
Declaration		I pronounce that *p* is the case, but *p* may not obtain.
		I pronounce that *p* is the case, but *p* does not obtain.

Table 2. Tentative taxonomy of Moorean speech acts

[77] Such assertions, though also arguably Moorean (i.e. satisfying the abovementioned criteria), did not get almost any attention in the literature; for a limited discussion, see Kriegel 2004, pp. 100-101. Cf. also Cappelen 2011 and McKinnon 2015, pp. 132-135, who argue from similar cases against doxastic norms of assertion.

3.1.2. Uniformity and completeness constraints

After doing this preliminary taxonomical work, let us now ask: should we expect a uniform explanation of the infelicity of *all* these constructions? In the present state, much of the literature on the topic seems to go in the opposite direction. Despite considerable attention in the linguistically oriented literature, the infelicity of Moore-paradoxical promises and commands were often provided with explanations pertaining solely to these types of speech acts or grammatic constructions by which they are often realized (e.g., Ninan 2005, Concoravdi and Lauer 2012, Van Roojen 2020, Mandelkern 2021) without the effort to generalize these results to other types of speech acts. But even more abstract and uniform theories that have this aim fall short of full generality. In *Speech Acts* (1970, p. 65) and further in the *Foundations of Illocutionary Logic* (Searle, Vanderveken 1985), Searle briefly suggests that the universal source of Moore-paradoxicality might just lie in the conflict between the expression of the sincerity condition of the speech act α and the content of the subsequent assertion. Given the above classification, however, this solution may only provide us with a straightforward explanation of the infelicity of omissive Sincerity pairs, for only such pairs contain the explicit denial of the sincerity condition. On the other side of the debate, some explanations focus on the *commitments* conventionally taken on by speakers (e.g. Woods 2018), treating Moore-paradoxicality as a conflict between what the speaker voluntarily commits to by α-ing and subsequently denying that they will fulfill such commitment. But unless such "commitment" is understood in a very broad way, such commitments encompass, at most, the speaker's actions[78], not their mental states, and hence can only explain the infelicity of Satisfaction, but not Sincerity formulations.

[78] Another problem for such commitment-based accounts is that in directives or declarations, the speaker does not take on any such commitments on themselves (and hence can't be expected to fulfill them), but (only in some cases) imposes them on the hearer.

All of these solutions, at least in their current state, remain partial – they either account for a peculiarity of only one type of speech act (promise, command, etc.), formulation (Sincerity or Satisfaction), or form (omissive or commissive). Though it is possible that a patchwork made of all, or some combination of these approaches would suffice to cover all the data we need to account for, it cannot be regarded as a satisfactory one, if our aim is obtaining a general explanation of the source of Moore-paradoxicality in speech.

The guiding intuition, according to which all of these constructions share a similar sort of absurdity ("Mooreanness"), demands a uniform explanation that is not currently on offer. But what if the intuition is wrong? The sort of linguistic phenomenology that prompts us into thinking that there is something commonly and distinctly wrong with such pairs was already extensively criticized in Chapter 1. But what we may do is apply the tests for Moore-paradoxicality presented there – PAST, THIRD, SUPPOSE – to find out whether there is some reason to regard them as such. Though these tests are explicitly formulated for assertions and beliefs only, the catalog of irrational attitudes and infelicitous speech acts was left open, awaiting precisely for this occasion.

To see how our sentence pairs fare on PAST and THIRD, let's compare our opening examples MP, MR, MA, and MC with their first-person past tense and third-person present tense versions (replacing, when necessary, an imperative or interrogative sentence with a declarative containing the explicit performative verb[79]):

(7) I promised I'll marry you, but I won't.

(8) He promised he'll marry you, but he won't.

[79] The role of explicit performative verbs in the "Mooreanness" intuition will be further discussed in 3.2.3 – *for now*, I will ignore this.

(9) I asked you to close the window, but I don't want you to do it.

(10) He asks you to close the window, but he doesn't want you to do it.

(11) I apologized, but I don't feel contrition for what I did.

(12) He apologizes, but he doesn't feel contrition for what he did.

(13) I ordered you to turn in your final paper by the end of the exam period, but you might not turn it in by then.

(14) He orders you to turn in your final paper by the end of the exam period, but you might not turn it in by then.

All of the mentioned constructions are, unlike MP, MR, MA, and MC, not typically infelicitous. Even more clear are the results of SUPPOSE:

(15) Suppose that: I promise I'll marry you, but I won't.

(16) Suppose that: I ask you to close the window, but I don't want you to do it.

(17) Suppose that: I apologize, but I don't feel contrition for what I did.

(18) Suppose that: I order you to turn in your final paper by the end of the exam period, but you might not turn it in by then.

In the absence of other indicators, we should think then of all constructions appearing in the table above (that is – at least those containing an explicit performative verb) as Moore-paradoxical and hence, demanding a unified explanation of their oddity. If one then sets out to provide a satisfactory account of Moore-paradoxicality in speech, then one ought to explain the infelicity of *all* Moorean speech acts, regardless of type, form, or formulation.

3.1.3. Continuity with the account for assertion and Stalnakerian pragmatics

Our goal in this chapter is to extend the account proposed in Chapter 2. to cover cases of Moore-paradoxicality which employ non-assertoric speech acts. To do so, it would be helpful to supplement the normative account with a framework that allows us to model the role different speech acts play in a conversation and how they affect other participants. As we have seen above, unlike assertion and other constatives, which can be in most cases assessed as proper or improper without any reference to its intended audience, many other speech acts essentially concern the impact of their performance on other disputants – and normative expectations towards the speaker ought to reflect that.

To extend my account in the desired direction, I will use Stalnaker's characterization of assertion based on his theory of dynamic pragmatics. Throughout many papers and books, Robert Stalnaker defended the model that spells the pragmatic effects in speech in terms of an interaction of the speaker's utterances with the *Common Ground* – the set of propositions accepted by the conversation participants. This idea was, at first, introduced to explicate the notion of presupposition (Stalnaker 1970, 1973, 1974): according to Stalnaker, we should understand it as something the speaker *does* rather than a property of a sentence or verb: for the speaker to presuppose that *p* is just for them to take *p* to be the part of the common ground. Later, Stalnaker extended on this idea to characterize what he calls "the essential effect" of assertion. According to him, the central aim of assertion is "to change the presuppositions of the participants in the conversation by adding the content of what is asserted to what is presupposed" (1978, p. 323). In this dynamic setting, the speech act of assertion not only expresses belief but aims at influencing the assumptions made by all conversation

participants; it ought to be understood as a proposal to alternate the common ground, unsuccessful only if the assertion gets rejected[80].

How should we understand the "common ground" in more precise terms? The quite standard way is to grasp it by appealing to the notion of common knowledge or common belief. The set of propositions in the common ground is then just understood as the set of propositions commonly believed or known by all conversation participants, that is known or believed by all, known/believed to all to be known/believed to all, known/believed to all to be known/believed to all to be known/believed to all, and so on *ad infinitum*[81]. Stalnaker later preferred (e.g., 2002) to characterize the common ground in terms of common "acceptance", not belief or knowledge, where "acceptance" is understood as a *sui generis* propositional attitude of which belief is a species. However, I side with Seth Yalcin (forthcoming; unpublished) in thinking that pure "common acceptance" is not enough either to fruitfully explicate the notion of the common ground or to explain the coordinated behavior in general[82].

---

[80] Alternatively, if one is less inclined to this interaction as a "proposal", one may simply say that the act of assertion adds the proposition to the common ground, and is later subtracted if the assertion gets rejected. But see section 3.4. for some reason for thinking in terms of proposals instead.

[81] Common knowledge or belief in this sense needs to be distinguished from the notions of "mutual knowledge" or "mutual belief" that *p,* which refer to situations in which all agents know or believe that *p*, but may fail to know/believe that others know/believe that *p*. Common knowledge or belief can be then characterized as mutual knowledge/belief of mutual knowledge/belief at every step of hierarchy. For more precise definitions, see Vanderschraaf, Sillari 2022; in Chapter 5. I will elaborate more on the justification of this assumption in explaining coordinated behavior.

[82] I will further discuss this issue in Chapter 5. But note also that a lot of the reasons for embracing "acceptance" as basic comes from the considerations not directly related to traditionally understood assertoric practice, but rather fictional or hypothetical discourse. If we employ the descriptive characterization of assertion as a conventional expression of belief used earlier, these discourses are

For present purposes, it would be sufficient to take common knowledge of what is accepted as a definition of the common ground and common knowledge as a "default setting" of the common ground[83].

Although the kind of normative analysis of assertion defended in the previous chapter tends to be pictured as an alternative to the Stalnakerian pragmatic analysis, this alternative is not an "either-or" choice. First and foremost, Stalnaker's apparatus was introduced only as a description of "the essential effect" an assertion has on the conversation, not an analysis of the speech act of assertion *per se*: Stalnaker explicitly states that his account is not a "definition", and treats "assertion" as a pretheoretically established notion (Stalnaker 1978, pp. 323-324; see also Clapp 2020, pp. 4-5). If one wants, as I do, to accept both the normative treatment of Moorean assertions and the Stalnaker-style analysis of the role of assertion in conversation, one need not choose one over the other. One may argue that these positions nicely supplement each other if we accept that the "default" setting of the common ground is common knowledge,

---

plausibly better understood as cases of belief pretense; see section 3.4. for further discussion of these cases.

[83] Another issue is that Stalnaker openly defines both assertoric and doxastic content (and hence – the common ground) in terms of his coarse-grained possible world theory of propositions. The problems with this theoretical move in connection to Moore-paradoxicality were discussed in Chapter 1. (for the overview of others see e.g. Clapp 2020, pp. 13-23). In 1.5.3. I suggested two ways to reinterpret my talk of "Moorean sentences" to "proposition-talk": by positing fine-grained propositional content or by relativizing belief and assertion in a proposition to a specific (linguistic) guise. I take the latter route to be more faithful to Stalnaker's original idea, which would require interpreting common ground as a set of propositions commonly known to be believed-under-a-guise, though it would lead to losses in the simplicity of intended modelling. Alternatively, one may also embrace the "impossible world" hyperintensional approach, and interpret propositions as sets of worlds sufficiently fine-grained for our purposes (Jago 2014) and retain the intuitive simplicity of Stalnaker's approach to context.

just as we considered the "default" empirical discourse to demand knowledge of assertors. As Yalcin (forthcoming) suggests:

"It has often been unclear how if at all [the] debate about the norm of assertion connects with the modeling proposal about assertion advanced [by Stalnaker]. But once we have in focus the question of what the default setting of common ground is, we can see a rather direct point of connection. (...) [I]f speakers as a default view the common ground as tracking common belief (knowledge), then unless the default is overridden, we expect them to assert only what they believe (know), to the extent they are cooperative." (Yalcin forthcoming, pp. 18-19)

Embracing this position fruitfully supplements the idea defended in the previous chapter that the normative standards for warranted assertion are determined relative to the expectations speakers have towards the epistemic prospects in a given discourse – in Stalnakerian terms, this could be just understood as an expectation of what kind of speakers' first-order attitudes feature in the common ground in a given discourse. On the other hand, if we take the idea that knowledge or rational belief is the norm of assertion to be basic (after Williamson's constitutive rule-view, for example), one may still quite naturally explain *why* assertions change the presuppositions of the conversation participants. Given that they treat the assertion made by the speaker as proper, they take it to express the speaker's knowledge or rational belief – and the fact that the speaker knows or rationally believes p by itself is a reason to also believe p (in the latter case – in the absence of evidential defeaters), due to the evidential support backing speaker's assertion.

Though Stalnaker himself did not theorize about how the force of non-constative speech acts could be captured by describing their interaction with the common ground, I take his model to be a natural setting for our desired theory due to its emphasis on the speaker-audience dynamics and explaining the sources of mutual normative expectations between them. In the next two sections, I will suggest a generalization of Stalnaker's treatment of assertion to other speech acts – first for

paradigmatic cases of commands and promises, and later for other directives and commissives, as well as expressives and declarations – that will elucidate the rise of Moorean infelicity in these other speech acts.

3.2. Moorean speech acts and the common ground

All needed elements are now in place. Let me then propose the central hypothesis of the chapter, based on two important elements of our diagnosis of what's wrong with Moorean assertions from the perspective of their interplay with the Stalnakerian common ground. It is fairly simple:

(Central Hypothesis) By performing the speech act α, conventionally expressing mental state *M* with a satisfaction condition *s*, the speaker *S* proposes to add to the common ground:

(B) the proposition that *s occurs* or s *will occur*.

(R) the proposition that *S is in M*.

In the case where α is an assertion, both conditions of CH express two features underlined in the connection between normative views of assertion and Stalnaker's framework sketched above. Since the speech act of assertion conventionally expresses belief and the satisfaction condition of such a belief is the truth of its content, condition (B) merely reiterates the Stalnakerian idea that the essential effect of assertion is to add the asserted proposition to the common ground. Concerning the (R) condition, it follows the general assumption that, if the assertion is accepted by the conversation participants and not outright called out as insincere or normatively defunct, the speaker is presumed to have made a proper *assertion*, i.e. it is assumed that the speaker believes what they asserted. Central Hypothesis, in short, proposes that *all* speech acts can be analyzed as having similar effects on the common ground with respect to the

kind of attitude they express. For example, by promising that they will *φ*, the speaker *S* is taken to propose to add to the common ground (B) the proposition that *S will φ* and (R) the proposition that *S intends to φ*; by commanding *H* to *φ*, the speaker *S* is taken to propose to add to the common ground (B) the proposition that *H will φ* and (R) the proposition that *S desires H to φ*; and so on. As CH explicitly mentions conventionally expressed mental state and such state's satisfaction condition, it can be, in its exact wording, applied only to constatives, directives, and commissives; in the next section, I will demonstrate how it can be extended to fit declarations and expressives as well.

The CH-based diagnosis of infelicity of our problematic speech acts is quite simple: performing them requires from the hearer to add contradictory propositions to the common ground or ascribe the speaker contradictory attitudes – in both cases proposing it exemplifies the irrationality of such a speaker. As an example, consider the following four types of Moorean promises as presented in Table 2.:

(P1) I promise to φ, but I don't intend to φ.

(P2) I promise to φ, but I intend not to φ.

(P3) I promise to φ, but I may not φ.

(P4) I promise to φ, but I will not φ.

Starting from Satisfaction forms (P3) and (P4), we proceed as follows: since the first conjunct is a promise – a commissive, conventionally expressing intention – we take it, along the lines of (B), to propose adding the satisfaction condition of expressed intention – namely, that the speaker will φ – to be added to the common ground. As the second conjunct of (P3) and (P4) is the assertion that the speaker will not or may not φ, we take it to analogously propose to add to the common ground the proposition that the speaker will not, in fact, φ, or that it is open whether the speaker will φ or not.

These two proposals are, however, incompatible, since adding the proposition that the speaker will φ to the common ground requires eliminating from it all possibilities in which the speaker will not φ, while in turn, both the proposition that it is open whether the speaker will φ or that they will not φ requires these possibilities to be included. Thus, we may think of the speaker as proposing to make two contradictory updates on the common ground.

For (P1) and (P2), we take, along the lines of (R), their first conjunct to add to the common ground the proposition that the speaker intends to φ. If we follow (R) in interpreting (P1) we may then, again, obtain a contradiction in the omissive case (for the first and the second conjunct aim at adding contradictory propositions about the speaker's intention to the common ground – first by (R), second by (B)). In the commissive case of (P2), we end up, by the same token, with the situation in which the speaker adds to the common ground both the proposition that they intend to φ and that they intend not to φ – which explains the intuition of the underlying irrationality. A similar, twofold approach works for directives as well: a speaker's command to φ, aiming at adding to the common ground the proposition that the hearer will φ and that they desire them to do so, gets contradicted by the subsequent assertion that they either will not (may not) φ or that they do not desire (or desire not to) φ.

CH's promise then is high when it comes to the problem we started with – finding a uniform treatment allowing for explaining the infelicity of all Moorean speech acts of all forms and formulations while maintaining continuity with the standard approach to the infelicity of Moorean assertions. In the following subsections, I shall justify the case for CH on more neutral grounds: firstly, I shall focus on the broad, "big-picture" justification of CH and then how it may account for a variety of otherwise puzzling linguistic data other than infelicity of Moorean conjunctions.

If we limit our attention to directives and commissives, we will find that CH's justification is well-entrenched in the discussion concerning the dynamic effects of

promises and commands. Though, to my knowledge, nobody defended a similarly straightforward analysis of the effects of directives and commissives in Stalnakerian terms, similar analyses were proposed. Portner's (2004, 2007) detailed treatment of imperative sentences goes in a similar direction, proposing that imperatives ought to be interpreted as attempted updates on a *To-Do List*, understood as sets of propositions ordered by preference relation. According to Portner, when the speaker orders the hearer to φ through an imperative sentence, one effectively aims at updating their preference relation in such a way, so that possible worlds in which the hearer φ's are preferred over those in which the hearer does not φ[84]. As Ninan (2005) observes, Portner's idea (if taken as an analysis of directive and commissive speech acts rather than imperative sentences alone) might be easily applied to infelicitous pairs such as:

(19) You must go to confession, but you're not going to.

(20) Shut the door! You're not going to shut the door.

(21) We're not going to go for a walk. Let's go for a walk.

In all of these cases, the imperative sentence's preference-updating effect is canceled by the assertion that eliminates worlds to be preferred from the space of commonly accepted possibilities. Mandelkern (2021) observes in response, however, that this approach cannot explain the infelicity of the MC-type pairs, in which the possibility that hearer will not φ is merely left open, not straightforwardly excluded (that is - *omissive* Satisfaction forms in my terminology). He concludes that a stronger mechanism governing the speech act of commanding is needed. CH might be naturally

[84] Following many others (e.g., Ninan 2005, Hacquard 2006), I simplify Portner's account for exposition: Portner originally takes imperatives' semantic content to be *properties*, not propositions, and *To-Do Lists* as preferences between worlds in which such property may be instantiated.

thought of as an explication of this idea, encompassing also other types of speech acts and associated conventionally expressed attitudes.

The idea standing behind strengthening the analysis of conversational effects of commissives and directives in the form of CH can be supported by thinking of those speech acts as linguistic tools allowing for rational coordination and common planning. Let me introduce the guiding idea here by an example. Suppose that we all want to throw a surprise party for our colleague, Jim, and to do so, we need to make up a workable, coordinated plan of action; the secret meeting at the office is called by our boss, Michael. We discuss our goals, and share relevant information about Jim and his daily routine: through subsequent assertions of the party planners, we come to commonly know that Jim usually gets back home after work around 7 p.m., that he hates chocolate, but loves coconut, and that he greatly admires fireworks. As the meeting enters the stage of final planning, Angela from the accounting utters:

(22) I promise to bring the coconut cake to the party.

As no one is however willing to go to the special shop selling fireworks, Michael decides to exercise his powers as a boss, turns to Oscar, and utters:

(23) I command you to bring the fireworks to the party.

It seems reasonable both for you and other meeting participants to take (22) and (23) as establishing a firm part of the plan they can coordinate around, if neither these speech acts were rejected or challenged. When Michael refers the party arrangements to Jim's wife, Pam, he may felicitously assert:

(24) Angela will bring the cake and Oscar will bring the fireworks.

(25) Jim will like the cake that Angela brings since it will be with coconut.

Furthermore, other party guests may rationally coordinate their actions – for example, arriving at Jim's house with a nice bottle of wine instead of sweets and

fireworks – around the information they gained after hearing Angela's promise and Michael's command. It seems that all of them can then *come to know* that Angela will bring the cake and Oscar will bring the fireworks by hearing these speech acts, which is what CH predicts, as both of these propositions are taken to be in the common ground. This would not be possible, if the only thing they came to know was that Angela now *prefers* bringing the cake to not bringing the cake or that Oscar prefers to buy the fireworks to not buying the fireworks (as Portner's "To-Do-List" analysis would imply). Furthermore, as evidenced by his felicitous assertion (25), Michael may integrate the knowledge he gained through previous assertion concerning Jim's culinary taste and through Angela's promise, which points towards the fact that one may treat her promise as having a similar effect on the common ground as ordinary assertion does. We may label the future-directed part of the common ground used as a basis for coordinated behavior the *Common Plan*.

If the above story concerning the role of promises and commands in coordination is true at least in its broad strokes, condition (B) becomes its natural consequence. What about (R)? At least two reasons for adopting it may be suggested. Firstly, if the "psychological" criteria for individuating speech types presented in 3.1.1 are correct, one may derive (R) quite straightforwardly by noting that to recognize how α contributes to the common plan, one needs to previously determine what type of speech act α is, that is whether α is a proper promise or a command. Determining this fact is, in turn, based (at least in part) on ascribing the relevant attitude being expressed. Thus, that the speaker is in a certain mental state is something that is added to the common ground in the same way in which it becomes the common ground that the speaker used English and performed a promise or a command unless later their sincerity is called into question. In short: while (B) specifies the intended primary effect α has on the common ground, (R) serves secondarily as a marker of the type of speech act α should be classified as, and that it is taken to be sincere.

Secondly, such expressed mental states often informally function as *reasons* as to why the relevant proposition is true or that something will occur. As I noted earlier, this is especially prominent when it comes to assertion: if it is normatively correct, it expresses rational belief or knowledge, and that the speaker has justification in favor of *p* may be used as a reason for introducing *p* into the common ground. But the similar is true for promises and commands: we oftentimes speak of acting on the basis of one's intention, or doing something *because* someone wants us to. Though the fact *S* desires *H* to φ cannot be literally accepted as a reason to believe that *H* will φ, another close connection holds: if the former fact is *salient* in a given context to *H* (and supplemented with additional factors, such as the social authority *S* has), it constitutes a reason for *H* to φ. If all conversation participants are trying to coordinate their actions, the fact that *S* made their desire salient through a command or request may lead us to expect that their desire will be satisfied[85].

Here I end with the philosophical, "big-picture" justification for adopting the Central Hypothesis and demonstration of how it explains the patterns of Moorean infelicity for promises and commands. Before focusing more closely on other types of commissives and directives, and extending the analysis to cover expressives and declarations, I will also demonstrate how CH may help us explain two otherwise puzzling pieces of linguistic data: the patterns concerning presupposition and challenges following the performance of promises and commands, and multiple

[85] This "intentionally making salient" feature also gives more plausibility to my initial claim that in conversational coordination contexts, contradictory desires or intentions strike us as odd (in commissive sincerity cases). While merely having contradictory desires or intentions is common and perhaps even natural, intentionally bringing them into salience and expecting others to treat our desires as reasons for action or believing that we will deliver on our intentions is irrational.

realizability of speech acts using declarative future-tensed sentences and attitude reports.

3.2.1. Supporting data: presupposition and challenges

The standard way to test our prediction, in line with Stalnaker's identification of propositions in the common ground with what is presupposed in a conversation, would be to find out whether it is felicitous for other participants of the conversation to presuppose the content supposedly added to the common ground. As mentioned already in Chapter 2., it is standard to observe that if the speaker asserts that *p* and this assertion is not rejected, it is felicitous for other conversation participants to presuppose that *p* is true and that the speaker *knows* or *believes* that *p*, as witnessed by the following four examples:

(26) A: It is raining now in Boston.

(a.) B: The Boston Marathon runners must be wet.

(b.) B: I hope it will stop soon.

(c.) B: How do you know that?

(d.) B: Why do you believe that?

Hence, the prediction is valid when it comes to assertion ((a) and (b) for the (B) condition, (c) and (d) for the (R) condition). What about promises and commands? If I am right, it should be felicitous for other conversation participants to presuppose that the speaker will φ and that they intend to φ if they promise to φ, and that they want their hearer to φ and that the hearer will indeed φ if they order the hearer to φ, assuming that the promise or the order are not rejected or challenged. These predictions are validated by conversational patterns exemplified below:

(27) A: I promise to bring both the cake and fireworks to the birthday party.

(a.) B: I hope it will be dairy-free.

(b.) B: Why do you intend to bring both? Can't D make the cake and you will buy the fireworks?

(c.) B: How do you intend to do it? The bakery and firework shop are pretty far from one another.

(28) A [to C]: I order you to bring the cake to the birthday party.

(a.) B: How many candles should I buy for the cake?

(b.) B: Why do you want C to bring the cake, if D makes them better?

Answers (27a.) and (28a.) are consistent with the (B)-based prediction for promises and commands, and answers (27b.), (27c.) and (28b.) – with the (R)-based prediction. The analogy between the assertoric and non-assertoric cases does not, however, stop here: one may also recognize that there is a similarity in how these speech acts are rejected or challenged. As one may reject the speaker's assertion by plainly denying what is asserted, another widely cited argument for strong doxastic norms of assertion (as mentioned in 2.3) observes that it is also common to challenge or reject assertions aggressively by questioning the speaker's epistemic authority or sincerity, e.g. by asking: "Do you really know/believe that?", or stating: "You don't really know/believe that!". Analogous things may be observed for promises and commands, for example in the following rejections and challenges:

(29) A: I promise to bring both the cake and candles to the birthday party.

(a.) B: You will not bring both!

(b.) B: Do you really intend to do it?

(c.) B: You don't really intend to do it!

(30) A [to C]: I order you to bring the cake to the birthday party.

(a.) C: I will not bring the cake.

(b.) B: Do you really want him to do it?

(c.) B: You don't really want him to do it!

This is, again, naturally accommodated by CH, as (a)-(c) challenges in both (29) and (30) either deny or question the (B) and (R)-related common ground proposals.

3.2.2. Supporting data: indirect performances and hedging

Another important feature supporting CH is an oftentimes overlooked fact that all speech acts (perhaps besides declarations) can be performed indirectly through multiple syntactic and lexical means. While most such performances involve the use of modal verbs ("should", "must", etc.) and/or using imperative or interrogative sentence type, it is also common to perform them *qua* future-tense declarative sentence or *qua* attitude report. The following examples may all be interpreted as constituting a promise:

(31) I promise to bring the cake to the birthday party.

(32) I will bring the cake to the birthday party.

(33) I intend to bring the cake to the birthday party.

and the following as constituting an order:

(34) I order you to bring the cake to the birthday party.

(35) You will bring the cake to the birthday party.

(36) I want you to bring the cake to the birthday party.

From the perspective of CH, the explanation for this fact is simple: (32) and (35) describe the intended (B)-effect for promises and orders respectively, while (33) and (36) the intended (R) condition; the speaker hence uses semi-equivalent means to achieve the same common ground goals. This does not mean, however, that we should

equate e.g., a promise to φ with an assertion that the speaker will φ or that they intend to φ, for the latter do not have, according to CH, equal double effect on the common ground. If the first conjunct is interpreted merely as an assertion about future events and not as a promise or a command, the following assertions still seem defensible:

(37) I will bring the cake to the party, but I don't intend to do it.

(38) You will bring the cake to the party, but I don't want you to do it.

as mere constatations of the fact that the speaker will perform an unintentional action in the future (37) or that the hearer will do something the speaker does not desire (38). Such predictive assertions will not play the functional role of promises and commands, as they are not accompanied by a presumption that the speaker intends or wants their hearer to bring about *s*[86].

The way in which a commissive or directive speech act may have an identical effect on the common ground as an assertion that *s* will occur may, however, be brought out by comparing promises and orders paired with the denial of knowledge or belief that *s* will occur, which show a similar infelicity to OMP and EOMP-assertions[87]:

[86] It is possible, as suggested by Grice (1971, p. 11), also to bring this difference using lexical means in English, using the verb "shall" or "should" instead of "will", which customarily emphasizes the intentional agency of the speaker or hearer. See also his discussion of presuppositions carried by intention avowals (1971, pp. 4-5).

[87] This point is also used by van Roojen (2020) to hypothesize that promises may themselves *be assertions*; I believe that CH, which postulates certain convergence between future-directed first-person assertions and promises in terms of their (B) effects on the common ground (yet distinct (R)-effects), can explain this intuition better without resulting in taxonomical collapse.

(39) I promise to bring the cake to the birthday party, but I don't believe that I will.

(40) I promise to bring the cake to the birthday party, but I don't know if I will.

(41) I order you to bring the cake to the birthday party, but I don't believe that you will.

(42) I order you to bring the cake to the birthday party, but I don't know if you will.

If we are right to model the common ground as what speakers take to be common knowledge in the context and assertion as guided primarily by the norm of knowledge, the infelicity of (39)-(42) can be easily explained, as adding, along the line of (B), the proposition that the speaker or hearer shall bring the cake to the common ground results in excluding the possibility that the speaker does not know that they will.

Notice also, that the same infelicity seems *not* to be exhibited by constructions containing attitude reports in place of the explicit performative:

(43) I intend to bring the cake to the birthday party, but I don't know if I will.

(44) I want you to bring the cake to the birthday party, but I don't know if you will.

This, again, may be explained along the lines of CH, as such attitude reports do not share either (B) or (R)-related effect with the respective assertion. Their status as alternative realizations of promises or commands may then be described as polite or hesitant, just as prefacing one's assertion with "I believe that" is understood as a *hedging* device. As noted in section 2.2.2., such *hedged* assertion paired with the knowledge denial is typically felicitous:

(45) I believe that it is raining now in Boston, but I don't know that.

If I am right in drawing this parallel, it would mean that promises and commands *qua* attitude reports should be interpreted as weaker, hedged ones just as assertions *qua* belief reports. Just as for assertion (which, if warranted, expresses at least *rational* belief), it means that in to properly promise to φ or order *H* to φ, the speaker needs to meet additional conditions apart from merely wanting *H* to φ or intending to φ (such as having relevant, socially recognized authority) that justifies adding the (B) proposition to the common ground. Nevertheless, the status of such hedged utterances as a weaker type of a promise or a command can be recognized, as they include descriptions of one of the intended effects on the common ground typical for these speech acts.

3.2.3. The role of explicit performative verbs

The case of multiple realizations offers an important insight into the nature of our Moorean examples, which I should comment on now. If performed *qua* future-tense declarative sentence or *qua* attitude report, our Moorean promises and commands come out either as plain contradictions ((46) and (48)), "epistemic contradictions" in Yalcin's sense (47) or contradictory attitude reports (49):

(46) I will bring the cake to the birthday party, but I will not.

(47) I will bring the cake to the birthday party, but I may not.

(48) I want you to bring the cake to the birthday party, but I don't want you to do it.

(49) I want you to bring the cake to the birthday party, but I want you to not do it.

While this aligns with CH's explanation of infelicity of all (employing explicit performative verbs or not) mentioned constructions, it also invites a reflection on the role of explicit performative verbs and the "Mooreanness" diagnostic introduced in

1.5.1. and applied in 3.1.2. The mentioned tests – PAST, THIRD, and SUPPOSE – were all applied to sentences containing performative verbs under the supposition that they are consistent. (46)-(49) don't meet this supposition; similarly, cases like Ninan's pairings of imperatives and denials of the satisfaction condition ((19)-(21)) or cases involving interrogatives (like MQ) are not even apt for these tests, since they do not have past tense or third-person counterparts or be meaningfully supposed.

In my opinion, this issue can be resolved if we introduce a distinction between these two classes – infelicitous constructions involving and not involving the use of explicit performatives – and think about what makes the constructions containing explicit performative verbs special. To do this, let us take a quick detour to the classic problem concerning the semantic contribution of explicit performative prefixes. It is as early as in Cohen's (1964) challenge to Austin's analysis of performatives where we see the problem of whether we should describe assertions of "I assert/claim/state that *p*" sentences and "pure" assertions of *p* as having distinct truth conditions. While the Austinian orthodoxy maintains that we should not do so, and think of "I α that/to/..." prefixes as only "making the illocutionary force explicit" (Austin 1962, p. 61; see also: Searle 1989, Jary 2007), many scholars pointed out that it is theoretically desirable to think of them as also contributing to *what is said* or *done* by speech act performance. According to this non-Austinian approach (e.g., Lemmon 1962, Ginet 1979, Bach, Harnish 1992, García-Carpintero 2013; cf. Williamson 2000, pp. 259-260), we may characterize α-ing with the use of such verbs as a form of *indirect* performance of α through related, self-verifying assertion that one is presently performing α. For example, according to such analysis, uttering the "I promise to φ" sentence constitutes a promise in virtue of this utterance being primarily an assertion that one is, indeed, promising to φ (i.e., in virtue of their assertion being true).

If this treatment of explicit performatives is true, we get another, only a bit more complicated explanation of why Moorean constructions employing such verbs (such as the ones present in Table 2.) elicit judgments of infelicity. Unlike in cases of direct

violations, such as (19)-(21) imperatives followed by the negation of the sincerity or satisfaction condition of a command, when the speaker uses an explicit performative, they violate CH-based normative constraints indirectly, by describing themselves as performing the speech act in question and subsequently stating that one of the specified effects of such performance is not met. In uttering the Moorean Promise, for example, the speaker *describes themselves as proposing to add* the proposition that they will φ by their first assertion and simultaneously proposes to add the contrary proposition by their second assertion. This, in turn, results in a similar problem as in the case of standard Moorean assertions discussed in the previous chapter: if the proposition that *S will not φ* is introduced to the common ground, *S*'s promise to φ cannot be performed successfully along the lines of CH. Since the latter is required for *S*'s *assertion that they promise to φ* to be true, the propriety of the whole utterance requires one to have a Moore-paradoxical belief of the form: *I believe that s occurs/will occur, but it does/will not* or *I believe that I am in M, but I am not*.

On this more fine-grained picture the cases discussed in this chapter should be distinguished into those directly violating CH-based normative constraints like imperative/interrogative-assertion pairs, which may be called "direct performative clashes", and "properly Moorean speech acts", interpreted as fully assertoric due to the role of the explicit performative verb. Of course, one needs not subscribe to this distinction, if one prefers the austere Austinian treatment. If performative prefixes only "make explicit" the illocutionary force of the utterance, as supporters of Austin would have it, all cases discussed here get a straightforward and uniform diagnosis of infelicity under CH and there is no need for further explanation. However, this move also comes at a cost of either inadequacy of the standard tests for Moore-paradoxicality[88], or abandoning the idea that cases discussed here are Moorean.

[88] Note also that, while it makes sense to consider (7)-(14) as past-tense or third-person versions of their source sentences, to maintain the principle of the test, the speech act performed with the

Though I will not adjudicate between these positions, I take the former, assertoric approach to be more in line with the general outlook adopted in this dissertation.

3.3. Extending the Central Hypothesis

Let me now turn my attention to how the proposed analysis could be adjusted to fit other, less obvious groups of speech acts: how should we interpret CH for other constatives, directives, and commissives, e.g. conjecturing, requesting, or threatening, and how should we adjust it to yield an analysis of expressives and declarations.

3.3.1. Extension 1: fine-tuning the analysis

So far, the proposed explanation covered only the paradigmatic cases of constatives, directives, and commissives: assertions, commands, and promises. Though not the strongest in their respective bundles, these speech acts are nevertheless "strong" in the sense that their performance is usually backed up by an additional authority of the speaker and “simple” in the sense of directly aiming at the introduction of its content into the common ground. One may wonder, then, how the proposed solution applies not only to those, but also to "weaker" and "complex" constatives, directives, and commissives: how, for example, one may account not only for the infelicity of Moorean commands, but also requests or questions; not only assertions, but also conjectures or suppositions; not only promises, but also offers.

---

utterance of (7)-(14) and their source sentences should be the same. If explicit performatives are assertions, this principle is maintained, as (7)-(14) may be felicitously asserted, unlike its first-person present tense counterparts.

The main difference to be applied in such cases is the specific content of a propositional attitude expressed that the specific type of a speech act expresses. Following our Bach-Harnish-inspired classification, we may say that while all constatives typically express belief, only flat-out assertion expresses belief *simpliciter*, and others express beliefs with the content directly tied to the illocutionary point of a respective constative. Conjectures then express not "weaker" belief in *p* than assertions do, but rather "the belief that there is reason, but not sufficient reason, to believe that *p*" (1979, p. 43), and suppositions "the belief that it is worth considering the consequences of *p*" (1979, p. 44); questions whether *p* express "the desire that [the hearer] tell [the speaker] whether or not *p*" and offers to φ "the intention to [φ] on condition that [the hearer] indicates he wants [the speaker] to [φ]"[89].

If an analysis along these lines is at least approximately adequate, then we should expect to observe the appropriate difference in the infelicity of Moore-like conjunctions depending on the type of speech act appearing in the first conjunct. While this is undoubtedly a pretty crude picture worthy of further refinement, when combined with CH it allows us to make correct predictions for which conjunctions should be considered infelicitous and which not. To see how it may be put to work, consider the four examples of conjectures paired with assertions that would in the standard fully assertoric case yield four Moorean assertions:

(C1) I conjecture that *p*, but ~*p*.

[89] Similarly, one may think of "stronger" conditions. If we follow, e.g., Turri (2013), in thinking that the speech act of *guaranteeing* normatively expresses one's reflective knowledge (i.e. when one guarantees that *p* one does not only express one's knowledge that *p*, but knowledge that one knows that *p*), or Turri (2015b) or Gaszczyk (2023) that speech act of *explaining* expresses one's understanding, we might expect different patterns of Moore-paradoxical constatives to emerge.

(C2) I conjecture that *p*, but it might be that ~*p*.

(C3) I conjecture that *p*, but I believe that ~*p*.

(C4) I conjecture that *p*, but I don't believe that *p*.

I take it that only (C1) and (C3) deserve the "#" label, while the omissive cases (C2) and (C4) are, given their standard reading, at least defensible. This fact is easily explained by the understanding of the speech act of conjecturing or hypothesizing along the lines of Bach and Harnish's analysis and CH. As the expressed attitude corresponding to the act of conjecturing is the belief that "there is reason (…) to believe that *p*" and this precise proposition gets added (by (B)) to the common ground, the question whether *p* ought to remain open for the participants of the discussion – hence (C1) is infelicitous, while (C2) is fine. Since, by (R), the proposition that the speaker has such belief is also added to the common ground, we see that having such belief rationally conflicts with the belief that ~*p* reported in the second conjunct of (C3), but not the lack of belief in *p* reported in (C4) – so, again, the commissive, but not omissive version is infelicitous.

A similar treatment might be developed or extended to cover other types of speech acts in the commissive and directive domains. One might consider, for example, cases of "weaker" directives than commands, e.g. requests, and observe that, analogously, uttering an omissive request-assertion pair:

(R1) Please, φ, but you might[90] not φ.

---

[90] R1 remains felicitous only if we consider it to consist of a request and assertion pair; one might feel the absurdity again if one interprets the "you might not φ" part as permission or a command to do the opposite of what gets requested in the first part. Given what was said before, the absurdity of such pairs is not particularly mysterious.

seems fine, while a stronger, commissive:

(R2) Please, φ, but you will not φ.

does not.

The same predictive principle works in the commissive domain, e.g. for threats, if they are taken to express the conditional intention to φ, provided that *H* does not ψ. All of the below constructions result in infelicity if uttered:

(T1) I threaten you that I will φ (if you don't ψ), but I might not φ (if you don't ψ).

(T2) I threaten you that I will φ (if you don't ψ), but I will not φ (if you don't ψ).

(T3) I threaten you that I will φ (if you don't ψ), but I don't intend to φ (if you don't ψ).

(T4) I threaten you that I will φ (if you don't ψ), but I intend not to φ (if you don't ψ).

but it does not for the following ones:

(T5) I threaten you that I will φ (if you don't ψ), but I might not φ (if you ψ).

(T6) I threaten you that I will φ (if you don't ψ), but I will not φ (if you ψ).

(T7) I threaten you that I will φ (if you don't ψ), but I don't intend to φ (if you ψ).

(T8) I threaten you that I will φ (if you don't ψ), but I intend not to φ (if you ψ).

CH, if paired with the standard assumptions about which mental states get conventionally expressed by such weaker or stronger speech acts, gives us accurate predictions and allows for adjusting to expected outcomes beyond paradigmatic assertions, commands, and promises to all cases in the three respective categories of speech acts.

3.3.2. Extension 2: adjusting to expressives and declarations

While Moorean commissives and directives may seem to be the "easy" cases, which the provided apparatus may straightforwardly explain, the question of expressives and declarations remains. Given the guiding principles of the taxonomy of speech acts employed here, both expressives and declarations lack one of the two crucial features employed in CH – either the mental state expressed by α is non-propositional and, as such, has no standardly associated satisfaction condition, or α does not, on the face of it, express any mental state at all.

How the proposed explanation can be adjusted to fit Moorean expressives and declarations with the same accuracy as Moorean constatives, commissives, and directives? We might do so by reflecting on what are the relevant propositions the speaker wishes to be taken for granted and form the basis of coordination between their hearers. In the case of expressives, it is natural to think that their intended effect is to let hearers know (or make them believe) that the speaker is in a relevant mental state[91] – they fulfill their conversational role just if they are sincere and are accepted as such. Hence, we would get the following version of CH tailored for expressives:

---

[91] This, of course, can be subject to some conventional norms of appropriate expression: "*Ouch!*" may be always sufficient to let others know that you are in pain, while truly showing that you regret your actions through an apology might require employing highly conventionalized means emphasizing

(**CH-Exp**) By performing the expressive speech act ε, conventionally expressing mental state *M*, the speaker *S* proposes to add the proposition that they are in *M* to the common ground.

For example, by apologizing for *φ*-ing, *S* proposes to add to the common ground the proposition that *S regrets φ-ing*. The treatment of Moorean expressives then fully resembles the explanation of infelicity we gave for Moorean commissives and directives in their Sincerity formulations: simultaneously adding the proposition that one is in *M* and that one is not in *M* or that one is in a mental state that excludes being in *M* is either impossible or irrational.

With declarations, the main problem in the wording of CH comes down to the fact that the content of the added proposition is tied to the propositional attitude conventionally expressed by the speech act in virtue of being its satisfaction condition. One may, I think, pretty easily sidestep this issue by noticing that declarations, as acts of conventionally making something the case, have something close to a satisfaction condition of their own – the fact the speaker aims to bring about. The primary intended effect of uttering a declaration on the audience is for the hearers to accept that some new fact has been created and take it for granted plainly in virtue of the declaration being uttered. Following this, we get a version of CH for declarations:

---

e.g. the intensity of regret. Saying "*Sorry!*" may be appropriate when you show your remorse over stamping on someone's foot, but would not do for the family of a victim of manslaughter.

(**CH-Dec**) By performing the declarative speech act δ, conventionally making it the case that *p*, the speaker *S* proposes to add the proposition that *p* to the common ground.

So, for example, by marrying *A* and *B*, *S* proposes to add to the common ground the proposition that *A and B are married*. CH-Dec, hence, allows us to account for the infelicity of MM-like constructions in both omissive and commissive forms: "δ, but ~*p*" and "δ, but it might be that ~*p*". Furthermore (in line with the observation from section 3.2.2.), it rightly predicts the infelicity of constructions in which the success condition of the declaration is embedded in the denial of knowledge or belief, as in:

(50) I pronounce you man and wife, but I don't believe that you are married.

(51) I pronounce you man and wife, but I don't know that you are married.

Nearly identical justification (from presupposition and challenge patterns and multiple realizability) as in 3.2.1 and 3.2.2. may be offered for such adjusted versions of CH for expressives and declarations. One thing to note is that, perhaps, while one may perform an expressive by self-ascribing the required mental state (e.g. apologize by uttering "I feel contrition for what I've done"), declarations remain highly conventionalized and do not usually allow such flexibility – saying "You are now married" would presumably *not* count as a proper performance of marrying, at least by the relevant granting authorities.

3.4. Objections

In this last section, I shall discuss a type of counterexample directed against the CH-based analysis, noted by Matthew Mandelkern (2021, pp. 12-13). Consider the

following scenario: Mark is getting kidnapped by some vicious-looking criminals. As he is carried away on the back of one of his assaulters, he cries out loud one of the following:

(52) I order you to let me go, but you will not!

(53) You might not let me go, but let me go!

These utterances still strike us as infelicitous, though it would be felicitous for Mark to cry out just "Let me go!" or "I order you to let me go!". But, given that it is commonly known between Mark and his kidnappers that they will disobey whatever he orders them to do and hence the satisfaction condition of his order cannot become part of the common knowledge, how is that possible?

This case brings Mandelkern to conclude that we should not understand commands as speech acts aimed at influencing the Stalnakerian common ground, but rather as acts embedded in a *conversational pretense* guided by the following norm (Mandelkern 2021, p. 14):

(*Posturing*) When you order someone to φ, you must act towards them as if you believe that they will φ.

While earlier I mentioned Mandelkern's view as one that applies specifically to one type of speech act, we can think now of possibly extending this picture further. In other work mentioned already in Chapter 2. (Mandelkern, Dorst 2022) Mandelkern endorses a similar view for assertion under the label of *Epistemic Posturing*; in principle, nothing also stands against formulating similar norms for commissive and expressive speech acts and declarations. Although delivering such an account would not be trivial, the real question it poses before us is this: wouldn't we be better off thinking of

CH as embedded in some kind of pretense rather than explicitly formulating it in terms of common knowledge?

Let me divide the answer to this objection into two: firstly, I will discuss how CH-theorist may accommodate Mandelkern's case, and secondly, I will consider Mandelkern's proposed solution on its own. Concerning the first issue, I think that Mandelkern is right that in certain cases we should think of the attitudes expressed by a speech act as embedded in a form of pretense. But if that is the main point, then it is hardly any news: many discourses require us to only pretend-believe certain propositions. In academic debates, we often presuppose that some proposition is true without believing it for argumentative reasons; in idealized inquiry, we tend to ignore certain complexities; during the trial, one may presume the innocence of the defendant to ensure fairness (Stalnaker 2002, p. 716). Most radically, we often knowingly engage in fictional discourse and discuss the fate of non-existent characters inhabiting non-existent worlds[92]. These cases are, however, easily accommodated by CH, if we simply adjust our understanding of the common ground to that of common knowledge of what the speakers *pretend-believe* in a given context. Given that it is public information available to all participants of the conversation, who know how to engage with such pretense beliefs, it is wise for them to adopt an "unofficial" common ground, which nevertheless obeys the same rules as an actual one, where the first-order attitude is treated as knowledge or rational belief.

This adjustment of Stalnaker's idea pairs well with the analysis of selfless speech acts hinted at in Chapter 2. As we saw in section 2.2.1., there is a tension between the intuitions of the infelicity of Moorean assertions and the felicity of Lackey's selfless assertions. I suggested that this tension can be resolved if we (borrowing the idea from

[92] Semeijn 2021, chapter 4. offers a detailed account of fictional discourse along the Stalnakerian lines.

Sosa 2011) take the selfless assertor to occupy a certain epistemic role in the context of their assertion. Notice that the general mold of Lackey's cases can be easily replicated in the non-assertoric context, as the two following scenarios modeled after CREATIONIST TEACHER show:

RAMBLING SERGEANT: *Sergeant Pepper lost control of his squad long ago. He does not believe that any of his words affect his subordinates. However, he regards his duty as a sergeant to include giving orders that are devised by his superiors. As a result, during the morning assembly, he commands: "I order you to shoot the prisoner", though he does not believe the command will have any effect.*

HESITATING SERGEANT: *Sergeant Pepper is a devout Quacker, who believes violence to be against God's will and sincerely wants to prevent it. However, he regards his duty as a sergeant to include giving orders that are devised by his superiors. As a result, during the morning assembly, he commands: "I order you to shoot the prisoner", though he does not want the command to have any effect.*

It seems that Sgt Pepper *can* properly order his squad to shoot the prisoner, even if his lack of control or religiously motivated desires are commonly known; yet it is still infelicitous[93] for him to say any of the following:

[93] Unless between the first and the second conjunct, the speaker intends to shift between the official and the unofficial common ground. Similarly for the case of fiction-telling: J. R. R. Tolkien *could* felicitously assert: "In a hole in the ground there lived a hobbit, but I don't believe that" if he intended the first conjunct to be uttered within the fictional pretense and the second to contribute to the official

(54) I order you to shoot the prisoner, but I don't want you to do it.

(55) I order you to shoot the prisoner, but I want you not to do it.

(56) I order you to shoot the prisoner, but you will not do it.

(57) I order you to shoot the prisoner, but you might not do it.

This gets, however, accommodated within the CH-based view if we think of Pepper as making an order from an idealized point of the role of a military sergeant, and what he takes the common ground to be in this situation.

But is Mandelkern's example a case of similar make-belief or role-playing? One could object to such reading, for the simple reason that, unlike in idealizations, debates, fiction-telling, or Sgt Pepper's context, what Mark said was intended to be *an actual,* not pretense-command, nor a command given from any specific role. Nevertheless, if Mark had any purpose in his speech act, I think it is then naturally interpreted as an attempt to change the common ground. These types of cases in part motivates Stalnaker's original description of a speech act of assertion (which I followed here) as a "proposal" to update the common ground rather than an outright update. Note, that while Mark's kidnappers could without doubt not only plainly reject, but easily challenge his authority in making the command by saying:

(58) You don't really believe that we will let you go.

(59) You know that we will not let you go.

these challenges only make sense if we assume that the standard that Mark's command is held to is that of what is commonly known or believed by Mark and his

common ground, though it would be still infelicitous if such context was uniform. Cf. also Wittgenstein 1998, §186 discussing a similar case.

kidnappers, not just whether it makes sense for Mark to pretend that he has some authority over his abusers.

Here we come to the second question of the extensional correctness of the *Posturing* (or similar) norm and its ability to account for the presented conversational patterns. Consider that one of the guiding ideas behind CH was that if assertions, promises, and commands are, in the default case, taken to add the respective satisfaction condition to the common ground interpreted as common knowledge, it helps us explain how such speech acts may be the source of knowledge (or at least rational belief) about current or future events and allow for success in coordination around such shared information. Can *Posturing* account for this phenomenon? Recall the party planning scenario I discussed at the beginning of section 3.2. If everyone, after hearing Angela's promise and Michael's order, became convinced only that Michael and Angela were *acting as if they believed* that Oscar will bring the fireworks and Angela will bring the cake, it would not make much sense to base their future course of action on the belief that the fireworks and the cake will be brought to the party[94]. Such conviction would make sense if the party planning was merely hypothetical (like Lewis' [1979, pp. 356-358] case of planning an imaginary heist) or were part of an act, that is – if everyone involved knew that the common ground involves pretend-beliefs. The presupposition and challenge examples (27)-(30) would make sense in such scenarios precisely because such contexts require everyone to play along the make-believe game and treat pretense-common ground as if it were real. But if standard cases were guided by pretense norms, (27)-(30) would exemplify only foolishness on the part of the speaker, who would fall for the act of the promisor or the

[94] A similar argument against Mandelkern and Dorst's account of assertion in terms of *Epistemic Posturing* is formulated by van Elswyk and Benton (2023, pp. 36-37).

commander. Mandelkern's case is then just not enough central example to change our interpretation of what should count as a default requirement towards speakers.

3.5. Conclusion

This chapter concludes the first part of this dissertation, the goal of which was to provide an explanation of Moorean infelicity in speech in terms of violating the norms guiding specific speech acts, which essentially require coherence at the level of belief. In response to the challenge concerning the discontinuity between the treatment of Moore-paradoxicality in assertion and other speech acts, I extended the treatment offered for Moorean assertions in Chapter 2. My proposal entrenched the normative considerations regarding assertion into a larger framework of Stalnaker's approach to pragmatics, by analyzing all speech acts in terms of attempted updates on the conversational common ground, interpreted as common knowledge of beliefs of conversation participants. Through sections 2. To 4., I defended this analysis in more detail, by providing its independent linguistic and philosophical justification (3.2.), and demonstrating how it might be adjusted to different specific speech act types (3.3.) and contexts involving conversational pretense (3.4.).

Nevertheless, some questions remained unresolved or only briefly touched upon. The first one, which I take to be too broad to fully answer in this chapter, concerned the role of explicit performative verbs in speech act performance. I suggested that the best account in terms of its continuity with theoretical assumptions made in this dissertation follows Bach and Harnish's analysis of explicit performative prefixes as serving a self-verifying assertoric role. This account essentially treats (despite the promise of the title of this chapter) Moore-paradoxicality as a phenomenon concerning exclusively assertoric speech, though to a much wider extent than it is usually presupposed. If one disagrees with me on this point, one is still

welcome to use the Central Hypothesis as an explanatory tool – though, in my opinion, it invites continuity problems for providing a uniform treatment of Moorean speech. Another question concerns the tacit assumption that common knowledge of what is accepted, understood as a layered structure of iterated knowledge states, is not only possible but fairly easy to obtain and widely employed. Some might rightly wonder whether such an assumption is justified. This issue, more broadly connected with the controversy surrounding so-called "introspection principles" in epistemology, will be directly addressed in Chapter 5.

Given the success of the strategy employed in these two chapters, we are now left with the important task of explaining the distinctive irrationality of Moorean thought. The next two chapters will deal specifically with outlining the overall approach that allows to demonstrate why Moorean beliefs are irrational (Chapter 4.) and defending its most controversial philosophical assumptions (Chapter 5.).

Part II: Moorean Thought

The whole intellectual life consists of beliefs, and of the passage from one belief to another by what is called "reasoning". Beliefs give knowledge and error; they are the vehicles of truth and falsehood. Psychology, theory of knowledge and metaphysics revolve about belief, and on the view we take of belief our philosophical outlook largely depends.

(Russell 1921, p. 231)

Chapter 4. Moore's Paradox in Belief and Doxastic Innocence

In the previous two chapters, I made my case for the explanation of Moore's Paradox in speech along the Priority thesis, which stated that Moore-paradoxicality in speech can be explained by appealing to irrationality in belief. In Chapter 2. I presented a case for the view that an accurate, discourse-sensitive normative account of assertion delivers the desired verdict, as it predicts that Moorean assertions are unwarranted in virtue of expressing beliefs that do not meet discourse-specific justification standards. In Chapter 3. I extended this case to other cases of Moore-paradoxicality in speech going beyond its paradigmatic cases in assertion. In both of these chapters, I simply *assumed* that Moorean belief is irrational – in this and the next chapter, I will try to answer the question of *why it is*.

The challenge standing before me is therefore to explain our intuition that Moorean beliefs are irrational, which, if I am right, is *the* central question posed by Moore's Paradox. In the philosophical discussion concerning this issue two competing explanations take the central stage: the Introspectionist strategy, which argues that Moore's Paradox should be seen as essentially a problem connected with the failure of self-knowledge (as defended in different variants e.g. by Hintikka 1962, Shoemaker 1995 and Smithies 2016), and the Self-Defeat strategy (as defended e.g. by Williams 1996, 1998, 2023), which locates the source of the problem in the fact that Moorean sentences cannot be simultaneously true and believed. The purpose of this chapter is to assess these two strategies with respect to their "innocence" – i.e. the strength of assumptions about the nature of rational belief made by these strategies – and the ability to explain the irrationality of *all* Moorean beliefs.

The sketch of this chapter is as follows: Section 1. will briefly discuss the wider issue of Moore-paradoxicality in thought and why Moore's Paradox in belief should be considered central. In section 2., I shall outline both approaches to Moore's Paradox

and explore their commitments when it comes to the characterization of the concept of rationality. I shall also state and clarify the mentioned metric of choice and preliminarily assess both strategies on this metric. In section 3. I shall present two families of doxastically problematic sentences: anti-expertise and iterated Moorean sentences and argue that given our characterization of Moore-paradoxicality, they should be regarded as Moorean in the wide sense. I shall demonstrate that this conclusion comes at a great cost for the self-defeat strategy, leaving the introspectionist approach unscathed. Lastly, in section 4., I shall take stock of these assessments and consider the prospects of providing an accurate, sufficiently generalizable, and innocent solution to Moore's Paradox.

4.1. The centrality of belief in Moorean thought

Before discussing possible explanations of the irrationality of Moorean belief, I should briefly comment on a wider issue of the presence of Moore-paradoxicality in thought in general. As noted in the opening chapter, Moorean belief is the most widely discussed case of Moore-paradoxicality in thought. But the same could be said of assertion and Moore-paradoxicality in speech – and as we saw in Chapter 3., this should not lead us to conclude that it obtains only in assertoric speech acts (at least traditionally understood). Can we then, in a similar vein, think of Moore-paradoxicality in thought as a phenomenon going beyond beliefs? In other words – *are there Moorean mental states other than beliefs*?

The first thing to clarify before answering this question is *what such Moorean mental states* would even look like. According to the tests for Moore-paradoxicality we employed, a Moorean *sentence* is characterized by either its asymmetric felicity in speech or rationality in thought. Moorean beliefs, in turn, may be simply assumed to be the ones that exhibit such asymmetric irrationality for some specific sentence. Following this characterization, these tests imply then that for a mental state ψ to be Moorean, it needs to be irrational to ψ that *p* if *p* is formulated in first-person present-

tense, but not otherwise, which in turn presupposes that ψ needs to be apt for evaluation as "rational" or "irrational" in the first place, as well as have propositional or sentential content.

The second condition leaves in the scope of our interest only propositional attitudes, which leads us to consideration of the traditional triad[95]: beliefs, desires, and intentions. There are, however, significant problems with considering the latter two attitudes as plausible candidates for being Moorean concerning the first of the mentioned conditions. For intention, the main issue is that they are essentially future-directed, and thus cannot exhibit the desired asymmetry concerning tense and allow for unproblematic iteration. It is perhaps strange, but by no means irrational to intend to do something unintentionally or while intending to do the opposite, just as to believe that one will falsely believe *p* at some point in the future. On the other hand, to say that one *intends now to intend now to* φ comes out as borderline nonsensical, unlike saying that one now believes that one now believes that *p*; and, as we saw, the potential for such occurrent iteration is essential for Moore-paradoxicality. As for desires, on the other hand, the problem is that they seem to evade any clear evaluation in terms of their rationality. In simple words, intuitively one can, without being labeled "irrational", desire *whatever*[96]. Even if one were to back out a bit from this claim (e.g. because they think that desiring that a flat-out contradiction is true makes one irrational), any characterization of norms of rational desire is prone to many questionable decisions which make it hard to claim that some desires deserve the

[95] However, we should remember also that both desires and intentions have their influential analyses in doxastic terms. For desires as beliefs, see Pettit 1987; for intentions as beliefs, see Velleman 1989.

[96] This thesis goes back at least to Hume's famous remark: "Tis as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than for the latter." (Hume 1739/1896, p. 416).

"Moorean" label[97], such as positing that desire distributes over conjunction. Neither intentions nor desires seem to be appropriate candidates for "Moorean" mental states as we understood it thus far.

Even if the approach that lists only beliefs, desires, and intentions as "basic" is a bit conservative, being more liberal does not change much. Imaginations, sometimes added to the list of basic propositional attitudes, just as desires do not obey rationality standards (and, *contra* Berkeley's master argument, it does not seem to be conceptually incoherent to imagine that *p* is true and one is not imagining that *p* [cf. Sorensen 1988]). A word should be perhaps said about the possibility of extending our repertoire to *factive* propositional attitudes, e.g. following Williamson (2000), who argues that knowledge should be considered a *sui generis* mental state. It is impossible to know that *p* and one does not presently know that *p* (though it is possible to know that *p* and that one didn't know that *p* or that someone else doesn't know that *p*) as the following proof easily shows (as in the rest of the chapter, *Kp* stands for *agent a knows that p* – I shall skip specification of the agent, as we will deal almost exclusively with single-agent deductions):

1. *K(p & ~Kp)* *(assumption for reductio)*
2. *Kp & K~Kp* *(1., distribution of knowledge over conjunction)*

---

[97] Wall 2012 is the only source of the intuition that there *are* distinct Moorean desires of the form: "I desire that (*p*, and I don't desire *p*)" or "I desire that (*p*, and I desire not *p*)". In Wall's view, such desires are irrational because their satisfaction globally results in one's desires being frustrated or one becoming indifferent to some state of affairs. I remain unconvinced by his case: it does not seem to me that a person who strives for indifference towards worldly affairs should be labeled irrational or fails to "grasp the concept of desire" (Wall 2012, p. 68). Such agents plausibly belong to a monastery, not a Logic 101 class.

3. Kp *(2., conjunction elimination)*
4. $K\sim Kp$ *(2., conjunction elimination)*
5. $\sim Kp$ *(4., factivity of knowledge)*
6. $\bot$ *(3., 5.)*

Note, however, that a similar proof could be provided for any factive mental state, provided that it distributes over conjunction. Similarly, one cannot currently regret that *p*, and that one does not regret (now) that *p*, if we take, as many do, "regret" to be factive. This is, however, not a particularly deep fact about the mental state of "regretting", but simply a semantic property of factive attitudes. Unlike in the case of Moorean belief, we are not dealing with irrationality, but with a straightforward impossibility – there is nothing particularly *paradoxical*[98] or irrational about "Moorean knowledge" or "Moorean regret", because, for rather unexciting reasons, such mental states simply do not exist. Another issue, of course, is whether this impossibility of "Moorean knowledge" is not the source of Moorean irrationality in belief – which will be one of the hypotheses I will assess in this chapter.

Given all this, I take beliefs to be the only plausible candidates for mental states that could be described as Moorean and which intuitive irrationality merits an explanation. As such, I will from now on set on the task of providing it, and regard Moorean belief as central in the explanation of all Moore-paradoxical phenomena.

[98] Though one may use this fact to derive a seemingly paradoxical conclusion, as in the standard derivation of the Fitch-Church Knowability Paradox.

4.2. Two anti-Moorean strategies and principles of choice

In this section, I shall present two approaches offered in the literature to argue that Moore-paradoxical beliefs are irrational: the introspectionist and the self-defeat strategies. After introducing them, I will focus on arbitrating between them and formulate and explicate a widely held desideratum of Doxastic Innocence, which a successful strategy needs to follow.

4.2.1. Introspectionist strategy

The first way of thinking about Moore's Paradox in belief locates the source of the unbelievability of Moore-paradoxical sentences in the properties of our rational introspective beliefs. The general idea behind this strategy is to think of Moorean beliefs as instances of failure of self-knowledge: it is natural to think of a believer who believes both that it is raining and they do not believe so to be somehow bizarrely inattentive to their own beliefs. As Sydney Shoemaker puts it:

"[Introspectionist] view, as you might expect, says that the content of such a [Moorean] sentence (...) cannot be believed without the subject believing a self-contradiction. The reason for this is that believing something commits one to believing that one believes it, in the sense that (…) if one believes something, and considers whether one does, one must, on pain of irrationality, believe that one believes it." (Shoemaker 1995, p. 214).

The thesis that if one believes that *p*, then one believes (or, more cautiously, one may rationally come to believe) that one believes that *p*, comes under the name of "self-intimation thesis", or *BB*. Combining it with another strong introspective thesis, "rational infallibility" (*4c,* stating that if one believes that one believes that *p*, then one believes that *p*), Declan Smithies (2016, pp. 407-408) arrives at the *rational biconditional thesis*, which he takes to be both supported in independently plausible picture of introspection and "explains why there is always some degree of irrationality

associated with believing Moorean conjunctions—or the conjuncts of Moorean conjunctions—of either omissive or commissive forms" (2016, p. 208).

Many similar explanations of the irrationality of believing Moore-paradoxical sentences follow from *rationalist* or *transparency* accounts of self-knowledge. Following Evans' (1982, p. 135) classic observation that we answer the question of whether we believe that *p* by employing the same resources as when we answer the question of whether *p*, some (e.g., Fernandez 2013) argued that knowledge of one's beliefs is just a part of what it is to be a rational agent and Moorean irrationality in belief is precisely a manifestation of such, using Shoemaker's phrase, irrational self-blindness.

To reconstruct the argument usually employed by Introspectionists, which demonstrates the impossibility of rational Moorean belief, let us make explicit three principles of rational belief they usually appeal to[99] (with "*Bp*" standing schematically for "agent *a* believes that *p*"):

(D) $Bp \rightarrow \sim B\sim p$ (*If one believes that p then one does not believe that ~p*)

(*BB*) $Bp \rightarrow BBp$ (*If one believes that p then one believes that one believes that p*)

[99] While the name "*BB*" is self-explanatory and widely used in the doxastic context, other names – *D* and *4c* – are borrowed straight from the conventional naming practice for axioms of modal logic they are doxastic counterparts of (originally *D* from "deontic", and *4c* for the converse of axiom *4*, that is, in our context, *BB*). Though as of yet we need not interpret these principles as modal principles, keeping this connection explicit will come in handy later in the chapter.

(*4c*) $BBp \rightarrow Bp$ (*If one believes that one believes that p then one believes that p*)

Assuming these three principles allows us to demonstrate that both OMP and CMP cannot be rationally believed (note that, unlike in the earlier proof for knowledge, we are not stating here that Moorean belief is impossible, but that it is irrational). Here is a proof for OMP:

1. *B(p & ~Bp)* *(assumption for reductio)*
2. *Bp & B~Bp* *(1., distribution of belief over conjunction)*
3. *Bp* *(2., conjunction elimination)*
4. *BBp* *(BB)*
5. *B~Bp* *(2., conjunction elimination)*
6. *~B~Bp* *(5., D)*
7. ⊥ *(5., 6.)*

and here for CMP:

1. *B(p & B~p)* *(assumption for reductio)*
2. *Bp & BB~p* *(1., distribution of belief over conjunction)*
3. *BB~p* *(2., conjunction elimination)*
4. *B~p* *(3., 4c)*
5. *~Bp* *(4., D)*
6. *Bp* *(2., conjunction elimination)*
7. ⊥ *(5., 6.)*

Though it is common to invoke *4c* to demonstrate the irrationality of believing CMP (see e.g. Sorensen 2000, Smithies 2016), we may easily provide a *reductio* of a commissive belief using only *BB* and *D*, making *4c* superfluous unless proven otherwise:

1.	*B(p & B~p)*	*(assumption for reductio)*
2.	*Bp & BB~p*	*(1., distribution of belief over conjunction)*
3.	*Bp*	*(2., conjunction elimination)*
4.	*BBp*	*(3., BB)*
5.	*~B~Bp*	*(4., D)*
6.	*B~Bp*	*(2., conjunction elimination, D)*
7.	⊥	*(5., 6.)*

This approach to Moore-paradoxicality in belief dates back at least to Hintikka's[100] (1962) seminal discussion. Repeating his argument, many have taken the ability of the introspectionist strategy to prove rational unbelievability (or, to use Hintikka's term: "doxastic indefensibility") of Moorean conjunctions to be an abductive reason to adopt the doxastic principles used in these proofs. The abductive argument remains, however, compelling only if the explanation it advances is actually "best" – and the introspectionist story is not the only one with a convincing claim for it.

[100] Hintikka provides only the first of the proofs and does not explicitly discuss the commissive version of the paradox. What is often overlooked, he also derives the "self-defeat" proof discussed below, though he does not endorse it as a correct account of the problematic nature of Moorean sentences (Hintikka 1962, pp. 70-71).

4.2.2. Self-defeat strategy

Another way to approach Moore's Paradox in belief is to observe that Moore-paradoxical sentences seem to have a lot in common with so-called pragmatic paradoxes (O'Connor 1948) such as:

(1) I am not using language now.
(2) Nobody ever uttered a sentence in English.

Sentences such as (1) and (2) bear a striking similarity to Moore-paradoxical ones in that they do have consistent truth conditions: it may well be true that nobody ever uttered a sentence in English, and certainly at times the intended referent of 'I' in (1) is not using language; however, again like Moore-paradoxical sentences, they seem to lack any appropriate performance conditions, i.e., they can never be felicitously asserted. It is standard to observe that it is the very act of uttering them that makes them false – in this sense, such sentences are pragmatically *self-defeating*. Of course, Moore-paradoxical sentences are not self-defeating in this limited sense, for they may be still true when uttered or thought; but one may easily extend this notion to belief and argue that Moore-paradoxical sentences are *doxastically self-defeating*, for they become false when they are *believed*. The standard argument for this conclusion in the omissive case may go as follows:

1.	$B(p \,\&\, \sim Bp) \,\&\, (p \,\&\, \sim Bp)$	*(assumption for reductio)*
2.	$Bp \,\&\, B\sim Bp \,\&\, p \,\&\, \sim Bp$	*(1., distribution of belief over conjunction)*
3.	$Bp \,\&\, \sim Bp$	*(2., conjunction elimination)*
4.	$\perp$	*(3.)*

and for the commissive case:

1.	*B(p & B~p) & (p & B~p)*	*(assumption for reductio)*
2.	*Bp & BB~p & p & B~p*	*(distribution of belief over conjunction)*
3.	*Bp & B~p*	*(conjunction elimination)*
4.	*Bp*	*(3., conjunction elimination)*
5.	*~B~p*	*(4., D)*
6.	*B~p*	*(3., conjunction elimination)*
7.	⊥	*(5., 6.)*

The most prominent defender of a pure version of this account is John N. Williams[101] (1996, 1998, 2023), who explains the irrationality of Moore-paradoxical beliefs in terms of their connection to truth rather than introspective consistency:

"As a guide to the truth, the omissive [Moorean] belief is worse than useless. Not only is there the absence of a watertight connection between the belief and its truth but there is a watertight connection between the belief and its falsehood. I am responsible in a special way for the falsehood of my belief, because what makes it false is not the content of my belief but my forming it. I have shot myself in the foot. Moreover, any evidence I have that warrants me in forming the belief—or continuing to have it—is evidence that warrants me in arriving inescapably at falsehood. The commissive belief is hardly any better as a guide to the truth. The belief escapes the fault of the omissive belief, namely self-falsification, only if I have overtly contradictory beliefs, and there is also a watertight connection between this pair of beliefs and falsehood, because one of them is bound to be false." (Williams 2023, p. 179)

[101] Other "self-defeat theorists" may include e.g., Deutscher 1967, Green 2007, and Sorensen 1988. It is worth noting that in his seminal discussion on pragmatic paradoxes, O'Connor lists Moorean sentences among them (see O'Connor 1948, p. 359).

The underlying logic of demonstrating *what's wrong* with Moorean beliefs along the "self-defeat" approach is very widely employed. Evidentialists, such as Adler (2002, p. 195) who take beliefs to be commitments to the truth of some proposition proportional to possessed evidence, accept the claim that one *ought to* dispose of any self-defeating or contradictory belief; likewise, those who claim that rational belief *aims at truth* (like Shah, Velleman 2005, but explicitly also Williams 2023, pp. 246-248) could endorse this solution. A very similar reasoning is employed by a variety of approaches to rational belief that take it to "aim at knowledge": if one thinks that we are rationally obliged to treat our beliefs as knowledge states, the self-defeat approach easily explains the irrationality of Moore-paradoxical beliefs given that knowledge entails both truth and belief (for a solution along these lines[102] see Huemer 2007, pp. 146-150; 2011)[103].

4.2.3. Doxastic innocence

Before trying to tie the explanation of Moore-paradoxicality in belief with any of these specific views – by finding a plausible account of introspective knowledge that justifies the use of *BB*, or arguing that rational belief should "aim" at truth or knowledge – we should ask first which of these general strategies of deriving the irrationality verdict we should adopt. As seen above, both strategies may have a claim to an explanation of why OMP and CMP are irrational to believe. Therefore, if one

---

[102] One might also interpret Williamson 2000 this way if one thinks that a similar solution should apply both to Moorean assertion and belief.

[103] The general idea is also quite pervasive (though not always explicitly endorsed) in the literature concerning transcendental arguments and pragmatic paradoxes in general: many likened Moorean sentences to Descartes' self-refuting "I do not exist" or Aristotle's response to Heraclitus' rejection of the Law of Non-Contradiction (for examples, see Haslanger 1992, p. 298, Stroud 2000, pp. 170-171).

wishes to argue that either one is accurate, one needs to not only prove that they have *a* solution, but that their preferred solution is in some way superior to the other. To be able to assess this theoretical superiority, we need to posit some choice principles, guiding our way to the solution of Moore's Paradox.

One such intuitive principle may be couched in terms of *how demanding* these strategies are when it comes to our ordinary notion of "rationality". Since holding Moore-paradoxical beliefs seems *overtly* irrational in a way that is not only discernible for infallible logicians but also for doxastic agents with limited cognitive resources – and the less we assume of rational belief, the better approximation of the source of our initial "absurdity" verdict we get. We may label this metric "doxastic innocence" and formulate it in the following desideratum:

(Innocence) The preferred solution to Moore's Paradox needs to make only *minimal assumptions* concerning the characteristics of rational belief.

How do our strategies fare on this metric? The argument taking Innocence as a more or less explicit premise had usually been employed to favor a self-defeat over introspectionist treatment of Moore's Paradox. Green and Williams (both self-defeat theorists) in their reviewal discussion of Moore's Paradox introduce the desideratum that "[i]f possible, an account of Moorean absurdity should not appeal to controversial principles" (Green, Williams 2007, p. 11), listing *BB* and *4c* employed by introspectionists among such controversial principles. In a similar vein, Williams appeals to self-defeat strategies' weaker doxastic commitments in his later work (see e.g., his 2015, pp. 30-32).

Though the argument seems intuitive, it is not easy to pinpoint how one should *minimally* characterize the self-defeat strategy in terms of what it demands from a rational believer, that is – how we should get *rational unbelievability* from *self-defeat*. As

mentioned above, proving that Moorean sentences are doxastically self-defeating requires accepting only that the rational belief distributes over conjunction and (for the commissive case) that one cannot rationally hold contradictory beliefs. But what does it take for rational believers to ensure that they *do not have such self-defeating beliefs*? Williams proposes that this question is answered by insisting that all rational believers are minimally required to follow the following norm:

> (*W*) Do not form—or continue to have—a specific belief that you can be reasonably expected to recognize is your very own self-falsifying belief. (Williams 2023, p. 374)

How should we answer the question of whether following *W* is a more doxastically innocent constraint on rationality than *BB*? I take it that an appropriate way to investigate this question, and strictly compare our two strategies on the Innocence metric, would be to use the tools of doxastic and epistemic logic. Though it is probably futile to expect such logical considerations to provide us an explanation of Moore's Paradox *per se,* as they posit highly idealized conditions on rationality (as, e.g., Sorensen 1988 pp. 19-22 rightly warns), I think it is still justified to use such tools to investigate the implicit commitments of the two strategies. For, obviously, by merely pointing out that ordinary agents may not rationally believe sentence *p* because of some or the other property of rational belief (expressed by one's preferred principle), we do not assume that rational agents are logically omniscient beings with unlimited computational resources, just that they may be expected not to hold *p* to the extent they are exercise their rationality. What I want to do is to show that, if some doxastic logic *L* adequately captures the commitments of one of the respective strategies, then if *L* admits rational belief in some Moorean sentence, real-life agents whose capacities are approximated by *L* could, in principle, rationally believe it as well.

Though there are many different ways to characterize a logic of belief[104], in what follows I will employ the most common way of doing so, by modeling the commitments of the two strategies in normal modal logic, where the sentential necessity operator (standardly denoted by "□") is interpreted as "agent *a* believes that...". Since I will deal only with single-agent systems, this operator will be simply written as "*B*", skipping the specification of the agent, as I did before. A *normal* propositional doxastic logic is characterized by a Distribution Axiom, *K*, and an inference rule called *Necessitation* (sometimes also known as the *Gödel rule*) specifying the behavior of the *B* operator:

(*K*) $B(p \rightarrow q) \rightarrow (Bp \rightarrow Bq)$ *(If one believes that p entails q, then if one believes p, then one believes q)*

(Nec) $\frac{p}{Bp}$ *(If p is a theorem of L, then Bp is a theorem of L)*

While there are also many possible semantic systems for such logics, I will stick with the standard relational semantics, spelled out in terms of Kripke's possible-world models. In the doxastic context, these systems of logic will be interpreted by classes of models characterizing the agent's doxastic state *(W, R, v),* where *W* is taken to be the set of possible worlds ("doxastic scenarios", as labeled by Holliday 2018), *R* – the relation of *doxastic accessibility* (the relation holding between the worlds w_i *and* w_j just in case if w_j is possible from a perspective of agent's beliefs at the world w_i) and *v* – a function assigning propositional variables to subsets of *W* (*v: Var* $\rightarrow$ *P(W)*), which establishes which possible worlds verify a given sentence (we say that $w \vDash p$ iff $w \in v(p)$). The agent is, in this modeling, said to believe *p* at *w* (that is – the sentence *Bp*

[104] As mentioned in 1.5.3. For an overview, see Lechniak 2011.

to be verified by w), when p is true in all worlds doxastically accessible from w ($w \vDash Bp$ if and only if $\forall u\ (wRu \rightarrow u \vDash p)$).

Our idea is then to interpret the principles of rational belief outlined above – *D*, *BB*, and *4c* – as modal axioms, which, when added to the basic normal modal logic *K* (characterized by *K* and *Nec*) give rise to stronger systems. From the semantic perspective, it is important to note that adding these axioms constrains the available properties of the accessibility relation *R*: while *K* is satisfied by models with an arbitrary *R*, adding *BB* (in standard modal logic referred to as axiom *4*) makes *R* transitive, *D* – serial and *4c* – dense. Other important axioms in our context include *T*, which corresponds to *R*'s reflexivity, and *5*, which corresponds with *R* being Euclidean:

(*T*) $Bp \rightarrow p$

(*5*) $\sim Bp \rightarrow B\sim Bp$

While *T* is not accepted as an axiom for belief, it is accepted in the logics of knowledge: if we would replace *B* with *K*, meaning "*a* knows that...", this axiom would express the fact that knowledge is a factive mental state. *5* on the other hand, often called the "negative introspection axiom", is rarely employed either for knowledge or rational belief apart from highly idealized contexts; as it plays virtually no role in our discussion, I will not assume it.

While it is quite obvious that a doxastic logic corresponding to introspectionist commitments is the logic *KD4*, a normal modal logic *K* with *D* and *BB* added as axioms or *K4!* (*KD4* supplemented with *4c*)[105], it is not immediately clear what logic would

[105] Name from Chellas 1980, p. 142. Adopting *K4c* in place of *KD4* (i.e., replacing *BB* with *4c*) does not allow to prove the irrationality of omissive Moorean beliefs.

correspond to the commitments of the self-defeat strategy. Although the base logic assumed in the proofs of the self-defeating nature of OMP and CMP in 3.2.2. is *KD*, what we are after is a logic characterizing *beliefs* of the agent who does not believe any self-defeating statement, not a logic that ensures the self-defeating character of Moorean belief *per se*. In a recent study, Adam Rieger (2015, p. 223) answers this question by proving that a minimal normal doxastic logic which delivers the verdict of unbelievability of *KD*-self-defeating sentences is a logic *K5c*, that is *K* supplemented with the axiom *5c* (the converse of *5*) called by him "negative belief infallibility"[106,107]:

(*5c*) $B{\sim}Bp \rightarrow {\sim}Bp$

The semantic constraint put on *R* by *5c* is a condition that states that for every possible world w, there must exist some doxastically accessible world v such that any world t accessible from v is also accessible from w (that is: $\forall w\, \exists v\, (wRv\, \&\, \forall t\, (vRt \rightarrow wRt))$). Though it does not have a specific name, it may be labeled "selective transitivity", since in an intuitive sense it states that for every world there is at least one "transitive" world doxastically accessible[108].

---

[106] The same axiom was shown by Linsky (1986) to allow for the doxastic version of Fitch's paradox to go through. For an illuminating discussion of minimal assumptions needed for obtaining Fitch's result in multimodal logics, importantly highlighting often-neglected differences between the Fitch-Church Knowability paradox and Moore's Paradox, see however San 2020.

[107] Interestingly, taken as a formula the axiom *5c* is just equivalent to *~B(p & ~Bp)* and may therefore even be alternatively labeled a "No Omissive Moorean Beliefs"-axiom.

[108] Note that, since the constraint on *R* is existential, such *R* is also serial, so any *5c*-logic is also a *D*-logic. Therefore, we write "*K5c*" instead of "*KD5c*".

The stance of the *knowledge-first* theorist sympathetic to Self-Defeat, who maintains that the irrationality of Moorean belief stems from the unknowability of its content, may also be expressed by adding a specific axiom of a normal bimodal logic of knowledge and belief[109]. To do this, let's assume minimally that the knowledge operator *K* obeys the *T* axiom, and that knowledge entails belief:

(*B*) $Kp \rightarrow Bp$

The desired axiom would then take the following form, closely mirroring that of *5c:*

(*B~K*) $B{\sim}Kp \rightarrow {\sim}Bp$

Since one cannot claim to *know* any doxastically self-defeating sentence, we may easily see why adding *B~K* delivers the same verdict as *K5c*. As *~Kp* is a theorem for any *KD*-self-defeating *p* (since knowledge is both doxastic, as expressed by *B*, and factive, as expressed by *T*) and belief obeys *Nec*, *B~Kp* is also a theorem; *B~K* then straightforwardly entails that one does not believe such self-defeating *p*[110]. I will refer to this bimodal logic simply as *B~K* logic.

This is all, perhaps, not as strong as the self-defeat theorist could have hoped for – after all, their initial promise was that Moore's Paradox is to be resolved without appealing to introspection-like principles. However, since $K5c \subset KD4$, we may safely say that the self-defeat strategy fares better than the introspectionist on the Innocence

---

[109] To obtain a semantic interpretation of such logic, we need simply to introduce to the model another relation, *R'*, interpreted as *epistemic accessibility.* Adopting axiom *B* entails that *R'* is an extension of the doxastic accessibility relation *R* (that is – all doxastic possibilities are epistemic possibilities) while adopting the *T* axiom – that *R'* is reflexive.

[110] For similar reasons, in *B~K* the belief operator would obey *the D* axiom, as contradictions cannot be known.

metric at least when it comes to OMP and CMP constructions. One may also defend *5c* (following Rieger 2015, pp. 222-225) as a standalone principle expressing the core idea of self-defeat strategy, that is that one must, on pain of irrationality, ensure that one eliminates false beliefs.

Innocence seems then to favor Self-Defeat, but it remains a preferred strategy only if it is extensionally equivalent to Introspectionism with respect to *all* Moorean sentences: what we have shown as of now is that they are equivalent in demonstrating the unbelievability of OMP and CMP. Recall, however, that focusing just on these two versions of Moorean sentences often led us astray; all previous chapters contained a plurality of examples that were arguably Moorean but did not follow the OMP/CMP syntactic mold. We might then rightly wonder whether these approaches are sufficient to provide a similar verdict to *all* Moorean beliefs, as characterized by the tests in 1.5.1., and whether any such sentences may be provably unbelievable given the *KD4* logic but not given *K5c* or its plausible (given the commitments and tactics of self-defeat strategy) extensions. In the upcoming section, I shall demonstrate that there are at least two types of such sentences – labeled in the literature "anti-expertise" sentences and versions of traditional, Moorean sentences with an iterated doxastic operator. Given that we are concerned with an explanation of Moorean irrationality at large, if a more innocent strategy cannot account for the unbelievability of *all* Moorean beliefs, it nevertheless should be rejected. Therefore, the following section should be understood as an argument against the self-defeat account.

4.3. More Moorean beliefs

In this section, I present two puzzling types of sentences: anti-expertise and iterated Moorean sentences, which will be shown to be border cases – they may be outlawed by the means available to the Introspectionist, but not the self-defeat theorist. The argumentative strategy goes as follows: firstly, I present the cases, then argue why

they ought to be classified as Moorean, and lastly demonstrate, why they cannot be plausibly accounted for by the self-defeat approach if it remains "doxastically innocent" in the sense outlined in the previous section, while remaining provably unbelievable given the assumptions of the Introspectionists.

4.3.1. Anti-expertise paradox

To introduce the first case, consider an example: imagine that you get approached by a mysterious cult leader, who offers you to join their newly formed religion. Though the leader seems persuasive and many sentences constituting the *Credo* of this religion strike you as relatable, you notice something peculiar. Among different propositions the leader tries to persuade you to believe, you find the following:

(3) God exists if and only if you don't believe that he exists.

Should you even consider the leader's proposal and join the group of people believing in (3)? Surely, many religious beliefs seem odd or even contradictory; in this sense, (3) is not an exception, or is, perhaps, slightly better, since such a god is easily conceivable (as a kind of Peter Pan's Tinkerbell *à rebours*). Yet, *believing* (3), either on religious grounds or not, is not odd in the same way as believing wildly improbable things – it seems deeply *irrational* or *incoherent*, not just evidently *incorrect*. But why?

Putting the content of (3) in a more formal outfit allows us to see that it is a case of what had become known in the literature as a statement of one's own "omissive anti-expertise"[111]:

[111] The anti-expertise sentences also bear a striking similarity to Burge's (1978) self-referential *Believer paradox*:

(OAE) $p \leftrightarrow \sim Bp$

which, with its close commissive counterpart:

(CAE) $p \leftrightarrow B\sim p$

was, by many (e.g., Sorensen 1987, Egan, Elga 2005) taken to be rationally unbelievable. But is anti-expertise Moorean and should, be covered by either self-defeat or introspectionist strategy?

In my view, there are at least three good reasons to think that it is. Starting with perhaps the weakest of the three, one may appeal to the idea dear to the heart of *knowledge-first* theorists – the fact that the *knowledge* version of omissive anti-expertise ($p \leftrightarrow \sim Kp$) is unknowable in the same way and even precisely for the very same reason

---

(S) I don't believe that *S* is true.

As I take the underlying logic of belief to be propositional modal logic, which does not have any quantifier axioms for belief nor an identity sign, I shall not be directly concerned with it here, as *S* does not have any straight formulation in the language of such logic of belief and knowledge. One may think that the primary lesson to be learned from *Believer* and its close kin *Knower* is that "belief" and "knowledge" should be interpreted as a sentential operator rather than a predicate in formal settings (as argued, e.g., by Montague 1963, Lenzen 1981). Surprisingly, J.N. Williams takes *S* to be Moore-paradoxical (1996), providing indirect evidence that he would classify anti-expertise sentences similarly. I comment more on this kinship in concluding remarks (6.2.).

as an epistemic Moore-paradoxical sentence. To see this, observe that $p \leftrightarrow \sim Kp$ is just equivalent to a disjunction of a knowledge version of OMP and CMP:

$(p \,\&\, \sim Kp) \vee (\sim p \,\&\, Kp)$

and that the second disjunct ($\sim p \,\&\, Kp$) is not only unknowable, but also plainly contradictory (as one may not know what is false). Hence:

1. $K((p \,\&\, \sim Kp) \vee (\sim p \,\&\, Kp))$ *(assumption for reductio)*
2. $\sim(\sim p \,\&\, Kp)$ *(PC, T)*
3. $K\sim(\sim p \,\&\, Kp)$ *(2., Nec)*
4. $K((((p \,\&\, \sim Kp) \vee (\sim p \,\&\, Kp)) \,\&\, \sim(\sim p \,\&\, Kp)) \rightarrow (p \,\&\, \sim Kp))$ *(knowledge of disjunctive syllogism)*
5. $K(p \,\&\, \sim Kp)$ *(1.,3.,4., K)*
6. $\perp$ *(5., proof in section 4.1.)*

As we may easily see, the *knowledge* version of anti-expertise is unknowable by the same token that the *knowledge* version of OMP is, as knowledge of one's omissive *epistemic* anti-expertise simply collapses to the knowledge of an omissive *epistemic* Moorean conjunction. As we shall see in a moment, this reasoning does not, however, transmit to the knowledge of one's *doxastic* anti-expertise.

The second reason to think that anti-expertise is Moorean is just observing that the abovementioned equivalence between a disjunction of omissive and commissive Moorean sentences and one's omissive or commissive anti-expertise holds. The intuition here is that on the most straightforward interpretation, we might only make sense of one's belief in (3) if such believer would hold a Moorean belief: either that God exists but they don't believe it or that God doesn't exist although they believe so. We should be, however, very careful about making too sweeping generalizations from this fact, for obviously simply believing a disjunction of *some* Moore-paradoxical sentences does not make one irrational. To use Van Fraasen's (2020) example: if I consciously do

not hold any belief about, say, the weather in Boston today, and believe (by the law of the excluded middle) that it is raining or not raining in Boston today, I will rationally believe a disjunction of Moorean omissive sentences: (*It is raining in Boston and I don't believe that) or (It is not raining in Boston and I don't believe that)*. Though not a direct proof then, this observation demonstrates one interesting and important fact: every omissive or commissive Moorean believer trivially believes in their own omissive and commissive anti-expertise.

The third and, in my opinion, most telling reason exploits our criteria for Moore-paradoxicality – the PAST, THIRD and SUPPOSE tests. For without any apparent problem, we might ascribe anti-expertise *to others*, including our *past or future selves*, as well as *suppose it* about ourselves. To introduce another example, imagine now that you and your friends go out to the pub after a long day of philosophical work; the pub hosts a trivia night with the topic being modern American cinema, and you all decide to form a team and join it. You start by relying on the opinions of your teammate, Mark, who voices them with apparent conviction. As it happens, it turns out that Mark's beliefs are extremely unreliable: after a few rounds, you still have zero points and haven't correctly answered any of the given questions. You decide to change the strategy – you go with every answer *opposite* to what Mark suggests. As you start to gain points and climb the team ranking, it is natural to suppose that you are well justified in believing that Mark is not only unreliable but *reliably unreliable* when the topic in question covers modern American cinema – that is, that he is a commissive anti-expert. Such an ascription does not put us in a logical predicament and seems as fine as an ascription of omissive or commissive error and, I take it, for similar reasons, it can be rational to believe in our past anti-expertise or suppose that we are anti-experts.

One might now object that our imaginary friend Mark has, in fact, similarly good grounds for ascribing anti-expertise to oneself and hence – believing himself to be an anti-expert. After all, the evidence we have for ascribing anti-expertise is also available

to him; if he may support his belief in his anti-expertise with such evidence, there may be nothing outright irrational in believing CAE (as claimed by Kroon 1990 or Richter 1990[112]). Suppose then that we all begin to discuss Mark's previous record on answering questions about American movies and he comes to believe in his anti-expertise. What then should Mark believe if we were to ask him whether, e.g., *Platoon* was awarded an Oscar for Best Picture in 1987? If his initial belief is that it wasn't, by recognizing his commissive anti-expertise he should in turn believe that it was. But given his thorough commitment to anti-expertise, he should then take his new belief as evidence for accepting that *Platoon* was not awarded an Oscar in 1987, coming full circle. By thoroughly believing in his own anti-expertise, Mark is then left in a logical predicament, unable to formulate a stable opinion about the matter, or even to consciously withhold such an opinion – for if he believes that he does not believe that *Platoon* was awarded an Oscar, he should then conclude (by *modus tollens* reasoning) that *Platoon* was *not* awarded an Oscar. Though one may follow Conee (1982) and insist that the rational thing to do for Mark would be to simply *stop* drawing such inferences (and, e.g., settle on his first change in judgment), denying that it is rational to follow *modus ponens* or *modus tollens* seems to be a harder pill to swallow than simply denying that Mark may rationally believe in his anti-expertise in the first place.

We may observe here that this strange asymmetry between the evidence admissible from the first and third-person perspective is not anyhow limited to the anti-expertise case. Such cases might be convincingly construed for Moore-paradoxical beliefs as well: if we consider the abovementioned scenario, we might easily imagine an expert on both American cinema and cognitive psychology, who informs Mark that:

---

[112] One may also read Christensen (2010, pp. 207-212) as eventually supporting this claim, though with important caveats.

(4) *Platoon* won an Oscar for Best Picture in 1987, but you believe it did not.

If we follow the idea that one may simply base their first-person beliefs on the same evidence as others use to form third-person beliefs, it would seem that Mark should simply believe such an expert and add a commissive Moorean sentence to his stock of beliefs – which is clearly problematic. As often assumed in the literature on belief change, such second-person announcements should lead one rather to update their beliefs only by the first conjunct of the announcement and change their introspective beliefs accordingly[113], believing, in effect, in their *past* rather than *present* commissive error[114]. Similarly, we may think of a second-person anti-expertise announcement as leading one not to believe in their anti-expertise, but to update their beliefs to accord with the converse of what they previously held to be true or to suspend judgment on the matter in which one's faculties lead them astray (as Egan and Elga suggest [2005, pp. 83-84]). In the form of a short slogan: the lesson to be learned from one's mistakes should not be that one is decisively mistaken, but that one should stop following the cognitive path that led to such mistakes. The same moral applies to the "standard" Moorean sentences and anti-expertise – which is another reason to think of them as close cousins in one broad family of paradoxical sentences.

Given that Moore-paradoxicality of CAE and OAE is established and hence any satisfying account of OMP and CMP beliefs' irrationality should extend to CAE and OAE beliefs, what do introspectionist and self-defeat strategies have on offer?

---

[113] See Van Benthem 2004 for a discussion of the problem in Public Announcements Logic; for different treatments of Moore-paradoxicality within the wider scope of logics of belief revision, see e.g. Lechniak 2018.

[114] As noted also by Hintikka 1962, pp. 90-91.

For the introspectionist, taking CAE and OAE to be irrational is in no way problematic. It is widely known (see, e.g., Lenzen 1981, Sorensen 1987 or, from a different perspective, Smullyan 1986) that *KD4* proves OAE and CAE to be rationally unbelievable. If, as the introspectionist claims, *BB* and *D* rightly capture the characteristic features of rational belief responsible for our judgments of Moorean irrationality, they may easily claim that CAE and OAE are Moore-paradoxical. Here is an exemplary proof for OAE (as *5c* is derivable in *KD4*, I will use it to make matters easier):

1.	$B(p \leftrightarrow \sim Bp)$	*(assumption for reductio)*
2.	$B(p \leftrightarrow \sim Bp) \rightarrow \sim Bp$	*(theorem of KD4)*
	2.1. $B(p \leftrightarrow \sim Bp)$	*(assumption for reductio of 2.)*
	2.2. Bp	*(assumption for reductio of 2.)*
	2.3. $B\sim Bp$	*(2.1., 2.2., K)*
	2.4. $\sim Bp$	*(2.3., 5c)*
	2.5. $\perp$	(2.2., 2.3.)
3.	$B(B(p \leftrightarrow \sim Bp) \rightarrow \sim Bp)$	*(2., Nec)*
4.	$BB(p \leftrightarrow \sim Bp)$	*(1., BB)*
5.	$B\sim Bp$	*(3., 4., K)*
6.	Bp	*(1., 5., K)*
7.	$\sim Bp$	*(5., 5c)*
8.	$\perp$	*(6., 7.)*

For the self-defeat theorist, however, such a move is unavailable. To see this, we may observe that OAE is consistently believed in the following *K5c* model:

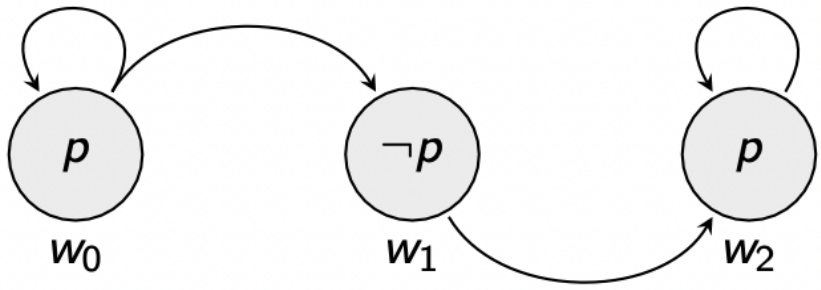

Figure 1. K5c model for OAE

$M_{OAE} = (W = \{w_0, w_1, w_2\}, R = \{(w_0, w_1), (w_1, w_2), (w_0, w_0), (w_2, w_2)\}, v = \{(p, \{w_0, w_2\})\})$.

while CAE in the following:

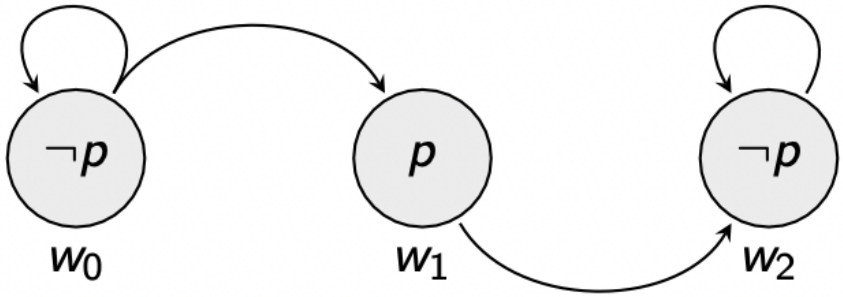

Figure 2. K5c model for CAE

$M_{CAE} = (W = \{w_0, w_1, w_2\}, R = \{(w_0, w_1), (w_1, w_2), (w_0, w_0), (w_2, w_2)\}, v = \{(p, \{w_1\})\})$

The fact that M_{OAE} and M_{CAE} are *K5c* models may be observed by inspection, as the doxastic accessibility relation *R* satisfies the "selective transitivity" condition mentioned in 3.2.3. Moreover, they are not only *K5c,* but also *K5c4c* (*K5c* + *4c*) models, since *R* is dense. Even more surprisingly, in both models, belief in OAE and CAE is

not only consistent but also not self-defeating, since in M_{OAE} $w_0 \vDash B(p \leftrightarrow \sim Bp) \,\&\, (p \leftrightarrow \sim Bp)$, while in M_{CAE} $w_0 \vDash B(p \leftrightarrow B\sim p) \,\&\, (p \leftrightarrow B\sim p)$. Therefore, even extending Self-Defeat strategy further by taking *K5c* or *K5c4c* as the base logic in self-defeat demonstration will not allow proving rational unbelievability of either commissive or omissive anti-expertise. For similar reasons[115], one cannot expect an appeal to *B~K* to help without strengthening the underlying assumptions about rational belief or knowledge.

By looking at the proof of OAE's unbelievability in *KD4* and at the properties of M_{OAE} countermodel, we might easily locate the problem of these weaker approaches. While all *K5c* agents (those who do not believe any self-defeating sentences) cannot consistently believe at the same time that *p* and that they are anti-experts concerning *p*, they can't *come to believe* that they do not believe that *p* when they believe OAE without becoming introspectively aware of their belief in OAE. This quite well captures the nature of Mark's predicament: being aware of the fact that he believes himself to be the anti-expert makes it impossible for him to set on the first-order belief or disbelief. The problematic *K5c*, *K5c4c,* or *B~K* agent is the one who believes in their anti-expertise but cannot introspect on this belief – which makes *BB* central for the proposed explanation of why one cannot come to believe OAE.

Can the self-defeat theorist appeal to any additional doxastic principles while retaining their contrastive innocence? Surely, *KD4* is not *strictly* the weakest doxastic or epistemic-doxastic logic that could prove the rational unbelievability of CAE and OAE; one might, simply, add a *no-anti-expertise-beliefs-axiom* (e.g., by stipulating that one is necessarily introspective only for anti-expertise beliefs or just outlawing OAE and CAE directly) and obtain a *strictly* weaker logic. But *KD4* is definitely the weakest such logic *in town* and such moves should count as *ad hoc*; moreover, they seem to have

[115] To see this, one may simply extend the *R* to form the reflexive relation of epistemic, instead of doxastic, accessibility in both models and observe that, in w_0, CAE or OAE are still not believed to be unknown.

no actual connection to the principles that the self-defeat or knowledge-norm theorists are advocating for. If anti-expertise is Moorean, then Self-Defeat cannot be said to provide a more doxastically innocent and equivalently general solution to Moore's Paradox.

4.3.2. Iterated Moorean beliefs

Let me move to the second case. Imagine now that you encounter an eager follower of our reverse-Tinkerbell religion, who asserts the following in your presence (example due to Sorensen 2000):

(5) There is a God, but I don't believe that I am a theist.

or another one, asserting in contrast:

(6) There is a God, but I believe that I am an atheist.

A natural reaction would be, I take it, again to consider such assertors irrational or incompetent users of the terms "theist" and "atheist". Surely, if they express their belief in God and understand that the term "theist" refers to those who believe in God's existence, they ought to recognize that they are one! There seems to be a comparable absurdity in believing (5) and (6) and holding the following Moore-paradoxical beliefs:

(7) There is a God, but I am not a theist.
(8) There is a God, but I am an atheist.

Given the standard understanding of "theist" and "atheist", (5) and (6) are equivalent to omissive and commissive Moorean sentences with an iterated belief operator:

(IOMP) *p & ~BBp*

(ICMP) *p & BB~p*

By the same token as with anti-expertise, I take it that the intuition about these cases is quite similar: they are also presumably weird and strike us as irrational, though it is fine to ascribe them to others and our past selves and suppose them.

How do these sentences behave when it comes to the comparison of self-defeat and introspectionist approaches? Interestingly, ICMP is provably unbelievable in *K5c* logic:

1.	*B(p & BB~p)*	*(assumption for reductio)*
2.	*Bp & BBB~p*	*(1., distribution of belief over conjunction)*
3.	*BBB~p*	*(2., conjunction elimination)*
4.	*BB~Bp*	*(3., belief in D)*
5.	*B~Bp*	*(4., belief in 5c)*
6.	*~Bp*	*(5., 5c)*
7.	*Bp*	*(2., conjunction elimination)*
8.	⊥	*(6.,7.)*

But the omissive version does not admit the same treatment and remains, just as CAE and OAE, not even self-defeating in the *K4c5c* logic, as shown by the following model:

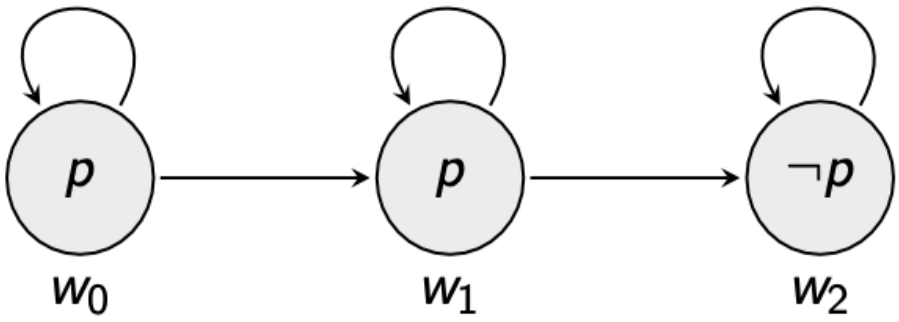

Figure 3. K5c model for IOMP

$M_{IOMP} = (W = \{w_0, w_1, w_2\}, R = \{(w_0, w_1), (w_0, w_0), (w_1, w_2), (w_1, w_1), (w_2, w_2)\}, v = \{(p, \{w_0, w_2\})\})$

In M_{IOMP}, $w_0 \vDash B(p \,\&\, \sim BBp) \,\&\, (p \,\&\, \sim BBp)$, which means that IOMP can be believed and is not self-defeating. Since *R* is reflexive, it satisfies the conditions for M_{IOMP} to be a model for *KT* logic, from which it follows that when interpreted as a doxastic model, M_{IOMP} is also a model for *K5c* and *K5c4c* doxastic logic. As one may easily see, however, IOMP could be easily proven to be rationally unbelievable, if *R* were transitive and serial – that is, if our logic of choice would include *BB* and *D*, as may be easily seen here:

1. $B(p \,\&\, \sim BBp)$ *(assumption for reductio)*
2. $Bp \,\&\, B\sim BBp$ *(1., distribution of belief over conjunction)*
3. Bp *(2., conjunction elimination)*
4. $BBBp$ *(3., BB)*
5. $B\sim BBp$ *(2., conjunction elimination)*
6. $\sim B\sim BBp$ *(5., D)*
7. $\perp$ *(5.,6.)*

Hence, the introspectionist may again easily account for believing a seemingly Moorean sentence where the self-defeat theorist cannot.

An objection that may come to mind here is that the introspectionist treatment of such iterated constructions generalizes easily to the unbelievability of bizarre and elaborate iterated constructions, which, supposedly, lose the apparent absurdity of (5). For the analogous proof of rational unbelievability as for (5) may be provided for any iterated construction, including:

($I^{10}OMP$) *p*, but I don't believe that I believe that I believe that I believe that I believe that I believe that I believe that I believe that I believe that I believe that *p*.

which may be rewritten as:

($I^{10}OMP$) $p \,\&\, \sim BBBBBBBBBBp$

or "$p \,\&\, \sim B^{10}p$" (where B^n means "*B* operator iterated *n* times") for short. Are *these* sentences (or even more baroque such as "$p \,\&\, \sim B^{10000000000000}p$") still Moorean? Sorensen (2000) reports the intuition that while such *omissive* complex, iterated Moorean constructions seem to decrease in absurdity as the iteration process goes up, the *commissive* constructions do not. The self-defeat theorist, remembering the fact that *K5c* invalidates the rational belief in iterated commissive sentences but not in the omissive ones, might embrace this result and argue that our initial response to (5) was misguided from the start.

This objection is, in a way, a reprise of a familiar objection to *BB* and other iteration principles, which holds that accepting them leads one into the uncomfortable position of accepting that whenever one believes that *p* one is thereby committed to believing $B^n p$ for any given *n*. The standard response here is to again clarify the methodological aim of upholding such principles and using the resources of doxastic and epistemic logic: by describing the beliefs of a perfectly thorough doxastic agent, we only provide an ideal to which we compare the performance of resource-bounded rational agents. Of course, one does not obtain infinitely many complex beliefs by merely believing an atomic sentence, but, in principle, one is in a position to deduce them from what they already believe. Though I^{10}OMP may not seem as *strikingly* irrational to believe as OMP or IOMP, the difference here lies merely in the computational complexity of the second conjunct rather than some deep fact about the concept of rationality. A way to hold this objection against such reply would be to posit the existence of some threshold of higher-order belief that is unavailable to the human mind (Williams and Green 2007, p. 22 defend this view, and point to Searle [1992, pp. 155-162] as inspiration). In this sense, you may, e.g., possess beliefs about your beliefs, but not necessarily beliefs containing a million iterations of belief operator. The epistemic rationality constraints would then require from you something that you are physically unable to do, and (by an epistemic variation on some sort of an *ought-implies-can* principle) this would demonstrate that they are wrong to begin with.

The issue here, however, obviously boils down to the chosen modeling. For starters, infinite belief iteration is not only a problem in *KD4* – for the *Nec* rule itself has an analogous layering effect, if the object sentence is a theorem of our logic[116]. As I indicated earlier, the results shown here can be interpreted in purely syntactic terms –

[116] This issue is distinct from the so-called "problem of logical omniscience", which will not be discussed here. For some helpful comments on the connection between logical omniscience and rationality, see Smithies 2015.

if so, then one may simply arbitrarily restrict the number of belief operator iterations that can be obtained in the logic's formulas and bound the axioms and inference rules accordingly. Nothing in the treatment of anti-expertise or iterated Moorean beliefs depends on such rules being unbounded. But it does not seem, on the other hand, clear to me whether the justification for such a restriction provided by the objector is actually persuasive. One may, I think, easily comprehend what is meant by $B^{1000000000}p$ as opposed to $B^{1000000001}p$ even if one is not able to adequately parse such an expression within one's lifetime, just as one may distinguish the sense of $s^n(0)$ from that of $s^{n+1}(0)$ in elementary arithmetic. To even conclude that while p is true, I don't believe^n that p (that is – to rationally believe an I^nOMP-sentence), I need to be able to comprehend such an iterated concept[117].

Taking stock from both of these examples, let us see how counting iterated and anti-expertise cases as Moorean impacts the considerations made on behalf of these two strategies. As we have seen, while the self-defeat approach fares better on the innocence metric if we think of Moorean beliefs as only the ones of the OMP or CMP form, it has significant trouble with providing the same verdict for our extended, yet

---

[117] Also, as Alan Hájek (2007, p. 224) points out, positing any precise threshold of comprehensible iterations of the belief operator seems to result in accepting a classic omissive Moorean sentence. For suppose such threshold is m; then one is committed to accept not only I^mOMP, but also OMP of the following form:

(OMP) I m^{th}-order-believe that p, but I don't believe that I m^{th}-order-believe that p.

since by believing that one believes that one m^{th}-order-believes that p one would commit oneself to comprehending $m+1$-order-beliefs.

sufficiently similar cases. The doxastic logic that forces the elimination of all *KD*-self-defeating beliefs – *K5c* – does not provide us with the verdict of the unbelievability of anti-expertise and iterated sentences. Furthermore, it does not (also when supplied by positive belief infallibility in the form of *4c* axiom) even provide us with grounds to prove that they are self-defeating, for these sentences remain possibly true when believed. Though one may come up with an alternative, still technically more "innocent" logic by simply introducing the negation of belief in anti-expertise as an axiom, or an axiom schema *5c*^{n}: $\forall n\ B{\sim}B^{n}p \rightarrow {\sim}Bp$ to get rid of iterated cases, this is an obvious *ad hoc* maneuver. The argument from innocence for the self-defeat strategy therefore stands defeated.

As for the introspectionist strategy, the discussion of these cases reveals one interesting fact: the right verdicts are delivered merely by *KD4*, which utilizes the positive introspection principle *BB*, with its converse, *4c* being unnecessary. Moreover, as we witnessed both in anti-expertise and iterated examples, adding *4c* without *BB* to our repertoire does not help us to dispose of any of them: hence, one should not see *4c* as anyhow useful in the discussion of Moorean cases (though one may, of course, argue for its truth on independent grounds). In short: in order not to be Moorean, one must be positively reflective, but not infallible.

4.4. Conclusion

In this chapter, I compared two approaches usually employed to demonstrate the irrationality of Moore-paradoxical beliefs: introspectionist strategy, which locates its source in the lack of self-knowledge, and self-defeat strategy, which locates it in the agent's commitment to false beliefs. I introduced an intuitive criterion of choice: Doxastic Innocence, which states that the preferred solution should make as weak assumptions about epistemic rationality as it needs to explain why Moorean sentences cannot be believed. I then demonstrated that two types of Moore-paradoxical (given theoretically neutral definitions) sentences – anti-expertise equivalences and iterated

Moorean conjunctions – cannot be shown to be doxastically inconsistent or self-refuting given the implicit commitments of the self-defeat strategy regarding the picture of rationality, while the introspectionist can account for their rational unbelievability by appealing to the "positive introspection" principle *BB*.

There are two conclusions one may draw from this chapter according to one's philosophical tastes: stronger and weaker. A weaker conclusion shows why and how a precise delineation of Moorean from non-Moorean cases impacts the claims made on behalf of different theoretical approaches. A theoretician who wishes to uphold a deflated explanation of Moorean irrationality not connected to self-knowledge owes us a precise argument on why we should not demand of them the explanation of irrationality of anti-expertise or iterated cases. As the tests for Moore-paradoxicality employed in this dissertation allow for counting them as Moorean alongside OMP and CMP, the self-defeat or the knowledge-norm theorist needs to show us a similarly neutral criterion disproving this connection.

A stronger conclusion, for which I advocated, is that the self-defeat (or any "self-knowledge light") approach to Moore-paradoxicality, as attractive as it seems, is mistaken. The possibility of providing a uniform treatment of OMP, CMP, anti-expertise, and iterated Moorean sentences (which bear important theoretical similarities) may count as a clear "best-explanation" argument for locating the source of their oddity in introspective failures rather than belief's necessary inaccuracy or falsity, as the self-defeat theorists would have it. Furthermore, their analysis also provides us with a better picture of which principles the introspectionist should be committed to. Although both *BB* and *4c* have their supporters, *4c* proves to be completely redundant in our endeavor, as it is both unnecessary and insufficient for proving the needed theorems. The strong conclusion of this paper then offers support to the picture of doxastic rationality as approximated by the *KD4* logic of belief and, more generally, to an introspectionist project of justifying epistemic iteration principles by appealing to our capacities as rational thinkers.

As noted already in this chapter, *BB* and other "introspection principles" are often considered controversial. Since I provided essentially only an abductive argument for adopting *BB*, this worry remains unresolved. The next chapter will deal with this issue by evaluating the most popular argument against adopting these principles posed by Timothy Williamson (1996, 2000). I will argue that even if one has the epistemic externalist intuitions represented by Williamson and his followers, one may find plausible grounds for defending some introspection principles, *BB* included, by siding with Alex Byrne's transparency account of self-knowledge. If I succeed, the solution to Moore's Paradox in belief (and hence, by extension – a full solution to Moore's Paradox itself) can be thus grounded in such an introspectionist approach.

Chapter 5. Knowing That One Believes Without Knowing That One Knows

In *The Principles of Psychology*, William James pronounces that "introspective observation is what we have to rely on first and foremost and always" (1890, p. 185). This unconditional trust in introspective evidence once seemed nearly ubiquitous in psychology and philosophy, but it is much less so nowadays. So-called introspection principles (among them *BB* discussed in the previous chapter), which may be seen as expressions of this epistemic optimism, have also fallen prey to this trend. Once widely assumed by philosophers and logicians across the board, proving useful in a variety of different formal endeavors, such as game theory and explanation of strategic behavior (Harsanyi 1967), doxastic and epistemic logic (Hintikka 1962), dynamic pragmatics (Stalnaker 1984) or artificial intelligence (Fagin et al. 2004), now are almost as commonly rejected by epistemologists[118] as descriptions of our cognitive powers and are, at best, taken to be true only of highly idealized epistemic agents.

While sociological reasons for this shift are complicated, it is fair to say that one of its most crucial causes was the rise of popularity of epistemic externalism and Timothy Williamson's forceful attack on the introspection principle for knowledge, *KK*, and his subsequent anti-luminosity position, the view that "nothing of interest is inherently accessible to [our knowledge]" (1996, p. 554) – which implies further the rejection of an introspection principle for belief, *KB*. Many epistemologists see Williamson's attacks against *KK* and *KB* as forming an "all-in-one" deal; some extend it even further (following Williamson [2014] himself) to reject *BB*. Typically, those

[118] For example, Louise Antony refers to *KK* and its "warranted belief" counterpart as "roundly rejected by epistemologists of almost every stripe" (2004, p. 12).

sympathetic to the externalist project and/or reliabilist analysis of knowledge follow Williamson in rejecting all principles expressing infallibility of self-knowledge, while those opposed to him, usually leaning towards internalist accounts of knowledge and justification, tend to defend (at least a qualified version of) *KK* alongside other introspection principles[119].

My aim in this chapter is to challenge this divide and offer a principled way of dissecting these two theses to arrive at (what I take to be) a moderate and broadly reliabilist position retaining much of the desired consequences of upholding certain introspection principles without, at the same time, defending *KK*. In the context of two previous chapters, this crucially involves the defense of two closely connected principles: *BB* and *KB*, expressing one's necessary *epistemic* access to one's beliefs. While my aim in defending *BB* is (given the content of Chapter 4.) self-explanatory, doing so for *KB* – a stronger principle, which entails *BB* – may not be. In section 1., I shall discuss more broadly the introspection principles and demonstrate how *the KB* principle may be crucial in allowing the possibility of *common knowledge of belief*, while at the same time avoiding the most common problems associated with *KK*. As the assumption of common knowledge of what is believed was tacitly assumed in the modeling accepted for Stalnaker's "common ground" framework in Chapter 3., this defense is necessary. In section 2. I shall argue that even fully accepting Williamson's anti-*KK* reasoning as presented in his (2000, ch. 5) and the externalist reasons guiding it, we may still reject the full "anti-luminosity package" and save the weaker

[119] An example of a somewhat isolated middle-ground internalist position might be Smithies (2019), who explicitly argues for the truth of an iteration principle for justification (*JJ*) without committing himself to defending *KK*. Greco (2014) argues from supposedly externalist premises for KK. See, however, Bird and Pettigrew (2021) for criticism of his and similar attempts; I concur with their conclusion that any plausible version of externalism entails the falsity of *KK*, though the converse does not hold (Bird, Pettigrew 2021, pp. 1731-1732).

introspection principles. In section 3., I shall argue that the conjunction of the theses *~KK* and *KB* is well motivated by transparency accounts of self-knowledge (most importantly Byrne's [2005, 2018]), and argue against previous attempts of grounding both *KK* and *KB* by appealing to these positions (section 4.); I shall also discuss possible threats to this view (sections 4. and 5.). If I succeed, then, given the success of the analyses of the previous chapters, I prove that even retaining broadly externalist intuitions about epistemic justification, one may justifiably employ *KB* in demonstrating the irrationality of Moorean beliefs (as discussed in Chapter 4.) and in justifying the common knowledge of belief assumption (used in Chapter 3.).

5.1. Introspection principles – overview and justification

The "introspection principles" discussed here are, on a more abstract level, principles guiding lower to higher-level interaction between epistemic or doxastic operators, such as "*a* knows that" (*K*) or "*a* believes that" (*B*) we already encountered in Chapter 4. They take us from the premise that the agent a^{120} is in some epistemic or doxastic state (e.g., knows that *p*) to the conclusion that one is thereby in a higher-order epistemic or doxastic state. These principles are not usually taken to describe the actual *performance* of epistemic agents, but rather their epistemic *competence* – therefore, the higher-order attitude is often worded with explicitly modal phrases, such as "being in a position to know" or "being justified in believing". I shall ignore these complications here, only reminding you that in our previous discussion, I explicitly defended the interpretation of *BB* as a principle of belief rationality, describing what one may *rationally come to believe* about their own beliefs, not an all-out accurate description of one's doxastic state (after all, Moorean belief is possible, but irrational). In this vein, I

[120] As in the previous chapter, I will skip the specification of the epistemic agent in question, unless it would be necessary.

will also assume that the agent in question possesses all relevant concepts, such as the concept of "knowledge" and "belief", and is able to perform the relevant deductions – the situation in which one gets struck by lightning in the middle of an introspective process will be safely ignored.

The two principles I will be most interested in here concern either iteration or interaction of knowledge and belief:

(*KK*) $Kp \rightarrow KKp$ (*If one knows that p, then one knows that one knows that p*).

(*KB*) $Bp \rightarrow KBp$ (*If one believes that p, then one knows that one believes that p*).

While *KK* and *BB* are most commonly referred to as "positive introspection principles" for knowledge and belief, to avoid confusion it is easier to address them specifically as "iteration principles" (after Greco 2015a), saving the "introspection" label for a broader application[121]. *KB*, sometimes also labeled "positive introspection" (Stalnaker 2006, p. 179), may be thought of as a plausible strengthening of *BB* since the truth of the latter principle is presumed to be grounded in the agent's infallible capacity for self-*knowledge*, obtained by introspective abilities; assuming that knowledge entails belief, *KB* entails *BB*. Although one may hold *KK* independently of introspection principles for belief, it seems intuitive to think of it as a further amplification of one's

[121] In an even wider sense, we may speak of "level-bridging" principles such as the "negative introspection principles" or "infallibility" principles (like *4c*; see Greco 2015a) – I will not discuss them in this chapter.

introspective powers[122]: according to *KK,* if one's belief amounts to knowledge, one is in a position to capture this fact *via* introspection.

As noted in the introduction, even though *KK* remains the main point of externalist attacks, all introspection principles are often assumed to be susceptible to similar pressure. An important point of departure here is of course Timothy Williamson, who explicitly rejects both *KK* and *KB* (Williamson 1996, 2000), as well as a version of *BB* for justified belief[123] (Williamson 2014). Yet, as we shall see, his arguments offered for this rejection are not monolithic and vary when it comes to their plausibility, especially when it comes to the scope of the famous "margin-for-error" principles. The guiding hypothesis of this paper is that one may occupy a broadly Williamsonian externalist position while at the same time rejecting some of these arguments: in particular, upholding *KB* (and hence – *BB*) and rejecting *KK*. Let's label this conjunction of views *~KK+KB* thesis.

Defenders of introspection principles (e.g., Greco 2015a) usually point out their two important applications: the treatment of epistemic-doxastic paradoxes and the widely made assumption of "common knowledge" needed to explain rational coordination between agents. This precisely mirrors two applications employed in the

---

[122] Unless one thinks that all "introspection principles" are not connected to any real-world introspective abilities possessed by actual agents, but rather are essential features of the chosen type of modeling of knowledge and belief (as seems to be suggested by Greco [2014] or Stalnaker [2006]) or as expressing conceptual truth about the *meaning* of "knowledge" and "belief", which seems to be the idea of Hintikka, who accepts both *KK* and *BB* as axioms of his logic, but not *KB* – and rejects the idea that such axioms should be justified by an appeal to introspective knowledge (1962, pp. 49-57). Such views merit their separate discussion which will not be pursued here.

[123] Important note: justified belief in Williamson's sense is a belief supported by one's evidence understood as one's *knowledge* (in line with his *E=K* thesis) and as such, Williamson's attack against the iteration principle for justified belief mirrors his second argument against *KK;* in a sense, it is more apt as an attack on the *BK* principle (see sections 4. and 5.) as understood in this chapter.

previous two chapters: demonstrating the irrationality of Moorean beliefs in Chapter 4. and explicating normative expectations of language users towards each other in "common ground" terms in Chapter 3., which made use of the common knowledge assumption. Concerning the first aim, as we clearly saw in Chapter 4., the treatment of the Moorean family of paradoxes requires adopting an iteration principle for *belief*, but not for *knowledge* – hence we need only defend *BB*, but not *KK*. Let me now examine, how and why introspection principles interact with the "common knowledge" assumption.

5.1.1. Introspection principles and common knowledge

At first, to grasp intuitively why a hierarchical layering of knowledge states may be essential for cooperative success, consider a toy example, in which Alice and Bob coordinate around where they will meet for a date. To establish the date, Alice sends Bob a text message saying: "I will be at Al's Cafe at noon" (p); let's also uncontroversially assume that she knows she will be there ($K_a p$). Thereby, she lets Bob know where she can meet her ($K_b p$). But is such mutual knowledge *enough* for them to coordinate and meet at the café expecting each other there, if they are both rational agents? No – because if Alice does not know that Bob knows p, she cannot rationally expect him to non-incidentally appear at Al's Café at noon, that is – to act in response to her message. To assume that, she also needs to know that Bob knows p ($K_a K_b p$). But if Bob takes Alice to be rational, then to expect that she will await him, he needs to make sure that she knows that he knows that she will be at the café ($K_b K_a K_b p$); and if Alice takes Bob to believe her to be rational, she needs to know that he knows that ($K_a K_b K_a K_b p$); and so on, and so forth. If we think of the Stalnakerian common ground in terms of publicly available information around which conversation participants may coordinate around (or, as I explicitly did, think of some parts of the common ground as establishing a common plan), this layering becomes essential also in running a successful conversation.

Of course, real-life, finite agents cannot be said to possess this sort of infinite knowledge; yet, it makes sense to say that they may safely assume it in many coordination scenarios. For this assumption to be justified, each of such knowledge layers should, at least in principle, be available to them. The appeal to introspection principles easily vindicates this idea. If Alice and Bob both know that *p*, they may come to commonly know that *p* by appealing to *KK* and the fact that they both assume their knowledge to be *symmetrical,* i.e. if Alice knows something, then *b* knows it and *vice versa*[124]. Since Alice knows that *p* then, by *KK*, she also knows that she knows it; then Alice may, simply by symmetry assumption, come to know that Bob knows it too (K_aK_bp). Since that is something that Bob may come to know by symmetry and *KK*, he may also come to know that Alice knows that Bob knows that *p* ($K_bK_aK_bp$) – this procedure may be repeated indefinitely, guaranteeing common knowledge of *p* (Greco 2015a).

The widespread rejection of *KK* led some to either voice skepticism about the possibility of so-construed common knowledge or argue that infinite iterative knowledge (*omega-knowledge*) is possible without *KK*'s help (Goldstein 2024, chapter 1). On the other hand, much less had been said on the matter of *common belief* or *common knowledge of beliefs*, perhaps because of the widespread assumption of the overreaching impact of criticism of *KK* on other introspection principles. This is nevertheless quite surprising, for to explain coordination success, we need not usually assume that what the agents believe is true, for we may perfectly well coordinate if we all believe something false (though with less success when it comes to outcomes of our

[124] Though in communication it is sometimes essential to make sure that the communication medium works properly with respect to the first knowledge-layer – e.g. for Alice to see the "read: 7:23 p.m." annotation below her message on WhatsApp – further symmetry is usually safely assumed in such contexts (e.g. it is not essential for Alice to see further "read..." annotations for the "read..." annotations).

coordination). In this vein, many philosophers write of "knowledge" in "common knowledge" as a concept "which can accommodate the mediaeval "common knowledge" that the Sun circles the Earth" (Heal 1978, p. 116) and treat the assumptions of common knowledge, common belief or common confidence as essentially interchangeable for all modeling intents and purposes (Lederman 2018). Assuming *BB*, common belief can be obtained in the usual way. If having a common belief is enough for us to coordinate, then *BB* alone saves much of the intuitive practical, and explanatory virtues of all introspection principles.

A fair point of criticism for this theoretical move was recently formulated by Yalcin (unpublished): it seems that pure common belief may not be sufficient to explain meaningful coordinational behavior. Yalcin provides the following example of "thoroughly Gettiered common belief" (p. 6):

> "In a noisy and darkly lit bar, Alice sees from behind a person she mistakes for Carl. (...) She leans in and whispers the following into the stranger's ear: "It's me, Alice. I've gotta run now, but let's meet tomorrow at the cafe at noon." The stranger is deaf, so he doesn't respond, but he gives a thumbs-up without looking at Alice. Alice interprets this as agreement about the plan, but in fact, the stranger is just signaling to the bartender that he wants another shot. Meanwhile, Carl is at another bar and coincidentally undergoes a completely analogous confusion: he mistakes a stranger for Alice, says "Let's meet tomorrow at the cafe at noon", and experiences the illusion of assent from the stranger. Alice and Carl leave their respective bars none the wiser, with both expecting to meet the other the next day at the cafe at noon." (Yalcin unpublished, p. 5)

Yalcin's intuition is that, although Alice and Carl commonly believe that they are meeting at the café at noon if they meet the next day, this will be a matter of coincidence not any meaningful coordination, just like Alice's waiting for Carl despite not knowing whether he got her message in our previous example. But does it mean that we need to insist on common knowledge instead of common belief? I take it to be an overreaction: the problem with Alice and Carl in the above scenario is not that they acted based on a first-order belief that they will meet, but that they didn't really *know*

that the other believed the same thing. If we choose to uphold *KB* instead of just *BB*, this verdict is naturally achieved by proceeding *via* symmetry assumption, since *KB*, in particular, entails the following principle[125]:

(K^nB) $Bp \rightarrow K^nBp$ (*If one believes that p, then, for any n, one knows*^{n} *that one believes that p*).

If we modify the scenario above so that Alice and Carl's common belief would be non-Gettiered, but instead assume that the café they are about to meet in bursts into flames at 11.50 a.m. (and hence they don't *know* that they will meet at the café), there is still a lot of rational, meaningful coordinational behavior going on that can and should be accounted for: Carl calling the café at 11 to make sure the tables are available or Alice leaving home at 11.30. These actions are rational because they both know that the other person *believes that p* (and that the other person knows that they know that and so on) and will thereby act on that belief. Similarly, in applications of this assumption to conversational dynamics, what is the common ground between the

---

[125] As any iteration *n* of *K* can be generated by the following procedure:

1. Bp *(assumption)*
2. KBp *(1., KB)*
3. $K(Bp \rightarrow KBp)$ *(KB, Nec)*
4. $KKBp$ *(2., 3., K)*
5. $KK(Bp \rightarrow KBp)$ *(3., Nec)*
6. ....
7. K^nBp

speakers need not be true to explain the communicational success, though plausibly still they need to *know*, not merely believe, that they have certain first-order beliefs[126].

There is then a lot to gain by defending *KB*, as it both proves why we are right in making theoretical assumptions regarding coordination used earlier in this piece and expresses a plausible justification for *BB* in terms of introspective abilities.

5.2. Williamson's anti-luminosity arguments

In this section, I shall focus on Williamson's arguments against introspection principles and find boundary conditions the *~KK+KB* defender needs to keep to stay faithful to the broadly externalist inclinations present in his epistemology. While Williamson's distaste for introspection principles is often monolithically derived from his broader "anti-luminosity" position and support for the safety conditions for knowledge, he does not always use the same argument when attacking them. Throughout his writings, he offers a variety of logically distinct arguments varying in their scope: while some of them target the general concept of a "luminous" condition – and both *KK* and *KB* as a result, others focus specifically on *KK*. Though this difference is oftentimes neglected, distinguishing at least two forms of the argument will be crucial for present purposes: the general argument against luminosity, utilizing the unrestricted form of the famous Margin for Error principle (as defended in [Williamson 1996] and chapter 4. of *Knowledge and Its Limits* [2000]) and a more specific anti-*KK* argument based the agent's knowledge of this principle (as presented canonically in [2000]'s chapter 5.)[127].

---

[126] As mentioned earlier, Yalcin himself (forthcoming) endorses a similar approach to defining Stalnakerian common ground.

[127] I am unsure whether Williamson himself makes this distinction. While in 2000, Williamson takes his anti-KK argument to rely on assumptions already "implicit in the argument against

5.2.1. Margin for Error Principles

The principle in question, central to Williamson's whole program in epistemology, is formulated as follows:

(*MAR*) If in case α_i one knows that *p*, then *p* is true in case α_{i+1}.

where α_{i+1} is a case sufficiently similar to α_i. Such cases may be understood as closely neighboring subsequent points of time (e.g., consecutive seconds or milliseconds), intensities of a given quality (e.g., minimal-degree increases in temperature), or worlds slightly differing in the size of a perceived object (e.g., 1-inch differences in height of a perceived tree); crucially, the difference between any α_i and α_{i+1} may be as small as one pleases, up to the point of absolute imperceptibility. In more formal terms, they can be understood as "close possible worlds": the intuitive sense of *MAR* is that for a specific belief to count as knowledge, it need not only be *true*, but *safely true*, that is true in all possible worlds which differ from the actual one, though this difference is, due to the limitations of our cognitive system, unrecognizable. In this sense, *MAR* can be seen as a generalization of the factivity condition for knowledge, which covers similar, that is imperceptibly different, cases.

Williamson's first formulations of the principle are used to motivate the epistemicist semantics for vague predicates (1990, pp. 104-106; 1994); however, later on, he came to hold that much (or all) of our knowledge is *inexact,* and to all such

luminosity" (2000, p. 17), he later tacitly acknowledges that *KB* and *KK* need not be offered a uniform treatment (2014, pp. 985-989) although he maintains that both are implausible (2014, p. 989). The two arguments discussed here are distinguished e.g. by Ramachandran 2012.

knowledge states, according to Williamson, *MAR* applies. Here is how he introduces the concept of inexact knowledge in *Vagueness*:

"The notion (...) is best introduced by examples. Vision gives knowledge about the height of a tree, hearing about the loudness of a noise, touch about the temperature of a surface, smell about the age of an egg, taste about the constituents of a drink. Memory gives knowledge about the length of a walk, testimony about the physical characteristics of a criminal. The list could of course be continued indefinitely. In each case, the knowledge is inexact. One sees roughly but not exactly how many books a room contains, for example: it is certainly more than two hundred and less than twenty thousand, but one does not know the exact number. Yet there need be no relevant vagueness in the number. The inexactness was in the knowledge, not in the object about which it was acquired." (Williamson 1994, pp. 216-217).

To see how *MAR* applies in those cases, consider the book case mentioned in this quote. If I am trying to tell how many books there are in a library room by looking, and my eyesight is not perfectly accurate, my belief that there are less than 1000 of them, when there are in fact exactly 999 would constitute merely "misplaced confidence", not knowledge. In Williamson's analysis, this is because my belief-forming mechanism (perception) is not sufficiently well correlated with the actual number of books for me to tell the difference between what is actually the case (α_{999}) and the case in which there is one more book present (α_{1000}). If, on the other hand, I form a belief that there are less than 2000 books in the room, this belief counts as knowledge, for it is true in similar cases, while false only in those in which the number of books is 2000 and more, which I would plausibly perceptively distinguish from the actual one and would not form such a belief. These sufficiently similar contexts mentioned in *MAR* in which *p* remains true form, in Williamson's words, a safety "buffer zone" (2000, p. 19) between the epistemic agent's cognitive success and failure. In this sense, *MAR* explicates a condition of "safety" (Sosa 1999) of knowledge and ensures its "reliability" (Goldman 1986), as present in the traditional externalist analyses of the concept (for more, see Ichikawa, Steup 2018). As said before,

Williamson in (2000) and later on endorses neither analysis and takes knowledge to be a primitive, *sui generis* mental state. While most do not follow Williamson on *that* claim, and neither do I, epistemic externalists who think that an analysis along reliabilist or sensitivist lines is correct largely agree on the plausibility of *MAR*.

5.2.2. General anti-luminosity argument

Assuming *MAR*, let us now look at the first, more general argument for the thesis that we may label General Anti-Luminosity offered by Williamson. According to this thesis, no non-trivial[128] condition *C* is *luminous,* in the following sense:

(Luminosity) For every case α, if in α one is in *C*, then in α one is in a position to know that one is in *C*.

Accordingly, Luminosity of the condition of "knowing that *p*" is equivalent to *KK*, and of "believing that *p*" equivalent to *KB* – General Anti-Luminosity entails the rejection of both principles. To set up Williamson's argument, we begin by constructing a sorites sequence of ever-so-slightly changing worldly conditions, modeled as a set of cases α_0, …, α_n. If we assume that in α_0 one is in a luminous condition *C,* we then proceed by utilizing characterization of luminosity and *MAR* to take us to the conclusion that one is in *C* in any case in the series. To construct the argument against *KK* along the lines of this general strategy, we need the following schema derived from *MAR*:

---

[128] In Williamson's terms, a trivial condition is one in "which one is always or never in" (Williamson 2000, p. 12).

"

(*K-MAR*) If in α_i one knows that *one knows that p*, then *one knows that p* in $\alpha_{i+1.}$

and against *KB*:

(*B-MAR*) If in α_i one knows that *one believes that p*, then *one believes that p* in $\alpha_{i+1.}$

If *K-MAR* and *B-MAR* are true, then knowledge and belief are not luminous, and hence *KK* and *KB* are false, by the following argument. Imagine Mr. Magoo looking at the modernist clock with just one hand and no numbers and dashes, and no way of knowing the time besides observing the face of the clock from a distance. By looking at the position of the hand, Mr. Magoo, despite his poor eyesight, nevertheless comes to believe something, e.g., that the time is between 3 and 5, or that it is *not* 6 o'clock; those beliefs are accurate and intuitively amount to knowledge (suppose it is 4 o'clock). Let the sequence of cases α_0, …, α_{120} be the sequence of imperceptibly different positions of a hand from 4 to 6; for simplicity, take the difference between any position to be 1 minute. Given that Magoo knows and believes that it is not 6 o'clock given its current position (α_0), and, by *KK* and *KB* – he knows that he does know and believe that. By *K-MAR* and *B-MAR* he would then also believe and know that it is not 6 o'clock if the hand of the clock were 1 minute closer to 6 (α_1); and here, as well, he would be in a position to know that he does. By applying the same reasoning 120 times, we arrive at the soritical conclusion: Magoo comes to know and believe that it is not 6 o'clock in $\alpha_{120,}$ that is when the position of the hand shows precisely 6 – which is, of course, a false conclusion. Since this result is unacceptable and the fact that Magoo knows in α_0 that it is not 6 o'clock is indisputable, *MAR* excludes the possibility of *KK* or *KB* being true. Neither knowledge nor belief is luminous.

The result Williamson arrives at here is very strong, given that it generalizes not only over beliefs but also over other mental states that are more commonly thought to be intimately connected with one's awareness – most prominently phenomenal mental states sometimes labeled "self-presenting" (Gertler 2012), such as pains or tickles[129]. What may seem initially suspicious about Williamson's argument is that it derives a highly general and strong conclusion concerning our introspective abilities without even mentioning what sort of process governs the acquisition of introspective knowledge. While intuitive examples used to justify *MAR* usually appeal (as I did two paragraphs above) to our limited abilities of perceptual discrimination, blurry memory, or impreciseness of someone else's testimony, making a similar case for introspection proves at least tricky[130]. *However we gain it*, introspective knowledge is accompanied by a kind and strength of authority unseen among other species of knowledge – if any knowledge is not inexact, the knowledge *that* one believes, perceives, or remembers something to be the case surely is (unlike knowledge *of* the fact one believes or remembers, or the object one perceives). Shouldn't knowledge gained through introspection be exempt from *MAR* until proven guilty?

---

[129] Some objections to General Anti-Luminosity limit their scope to defending the luminosity of only these states, e.g., Weatherson 2004 and Conee 2005.

[130] The most basic problem of applying *MAR* to introspective knowledge comes right at the start: between which possible mental states one is to be unable to discriminate? One may perhaps think here of a series of different mental states differing in their intensity (in the case of feelings) or, in the case of belief, subjective confidence in the truth of a proposition (for a similar case see: Silins 2012). But this seems problematic only if we maintain a combination of a strong descriptive Lockean thesis (according to which one's belief in *p* is just equivalent to having confidence in *p* above some threshold) and the more plausible view that one is unable to ascribe to oneself a precise level of confidence. Since there are good reasons to reject the descriptive Lockean thesis, I take this sort of objection to be unpersuasive. At the very least, the burden of constructing a plausible case of this type lies on the proponents of *MAR*'s application to introspective knowledge.

Many have taken issue with this lack of distinction between different sources of knowledge and proposed that *MAR* should be limited to certain source-indexed knowledge operators, like *knowledge-by-perception*, *knowledge-by-memory*, and so on, while, e.g. *inferential* or *introspective* knowledge should be free of it (e.g., Dokic, Égré 2009). This proposal is problematic since it severely limits our ability to model the interplay between knowledge coming from different sources (and there are convincing arguments showing that many of them also sometimes need a margin for error, see, e.g.: Lasonen-Aarnio 2008 for an argument concerning some cases of *inferential knowledge*); but exempting self-knowledge from *MAR* does not require such radical moves. To make introspective knowledge provisionally immune to *MAR* without rejecting the principle altogether, we may simply assume that the syntactic form of *p* in *MAR* is limited to that of a sentence free from epistemic or doxastic operators[131]. We may then rightly wonder: what happens to *KK* and *KB* if we accept only this restricted version of *MAR*?

5.2.3. Restricted anti-*KK* argument

To answer this question, let us turn our attention to the second version of Williamson's anti-*KK* argument (as featured in 2000, chapter 5.). Unlike in the first argument, however, let us now stipulate that although Mr. Magoo's higher-order knowledge is not sensitive to *MAR*, he also *knows* that such an amended version of *MAR* is true of his first-order knowledge. If Mr. Magoo knows that he knows that it is

[131] This crude limitation comes with some more or less obvious caveats. Most importantly, the language used needs to disallow the possibility of encoding higher-order beliefs into first-order ones; it will be sufficient for our purposes to take our model to be that of formulas of propositional multimodal logic as presented in the previous chapter. One may also worry about second-order self-knowledge gained through non-introspective means (e.g. through testimony of one's therapist); such cases, problematic in their own right (see section 6.2.), will not be dealt with here.

not 6 o'clock in α_i, by his knowledge of *MAR* he may still come to know that it is not 6 o'clock in α_{i+1}, for he knows that whenever he knows that it is not 6 o'clock in α_k for some *k*, it remains true in α_{k+1}. This, in turn, validates *K-MAR* as applied to knowledge of one's first-order knowledge (assuming additionally one-premise closure) without assuming anything of substance about one's introspective abilities – and we may again derive a paradoxical conclusion. In effect, to reject *KK* via the second route, we trade the unrestricted application of *MAR* for an additional assumption concerning one's knowledge of *MAR*. The same reasoning cannot be, unlike the previous argument, plausibly applied to *KB*, as first-order belief does not come with a margin for error, nor one needs to believe of their beliefs that it does[132].

There is then clearly room for rejecting *KK* for broadly Williamsonian reasons while maintaining support for *KB*: such a position needs only to take *MAR* to apply to one's first-, but not higher-order knowledge. I shall label the combination of these views the *~KK+KB* thesis. Whether it is a good position to occupy comes down to how we characterize our introspective capacities: whether self-knowledge of beliefs is sufficiently exact to justify abandoning MAR or whether we are otherwise obliged to take our beliefs to yield to *MAR* principle just as we are obliged (by the logic of the above argument) to treat our knowledge.

The challenge standing before us is therefore this: we need to find a theory of self-knowledge that both (a) vindicates the idea that higher-order knowledge of one's beliefs is not subject to *MAR* and at the same time (b) stays faithful to the externalist intuitions about knowledge motivating restricted *MAR*'s application. The following sections will discuss the prospects, limitations, and challenges of exactly such a project.

---

[132] For more on this point see sections 4. and 5.

5.3. Transparency of belief and *KB*

James follows up his statement of epistemic optimism which opened this chapter with a no-less optimistic characterization of introspection: "[t]he word introspection need hardly be defined; it means, of course, the looking into our own minds and reporting what we there discover." (James 1890, p. 185). Unlike his optimism, the vision of self-knowledge presented here is more prevalent and widely represented. Starting from Locke and Hume, many philosophers characterized introspection as an "inner sense" or "inward perception", and knowledge of one's mental states as essentially perceptual – most notable contemporary examples include Armstrong 1968, Lycan 1996, Nichols, Stich 2003. This vision of introspection also seems to be essentially the one endorsed by Williamson, e.g. when he justifies the intuitive rejection of *KB* by noting that it is "at least as hard to introspect the time according to [one's] phenomenal clock as to see the time according to a real clock" (2014, p. 989), presupposing that we come to know what we believe by looking at how things appear to us – presumably, with a "mind's eye".

It is easy to see why this account of introspection makes the application of *MAR* to self-knowledge compelling. For if knowing that one is in mental state *M* comes from one *perceiving M*, then no matter whether one does it by "looking into one's own mind", as James would have it, or *via* Armstrong's "self-scanning" procedure, this perception plausibly has to admit *some* possibility of error. As exempting self-knowledge from *MAR* would require not only assuming the existence of additional perceptual mechanism, but an infallible one at that, I believe this approach to introspection should be abandoned if we are set on justifying *KB*.

Among contenders to "inner sense" theories, we might crudely distinguish at least two of their bundles: *acquaintance theories*, building upon Bertrand Russell's remarks on acquaintance with the mental (1912), and *transparency theories*, building on

Gareth Evans' work (1982)[133]. As the former's stronghold field of application is knowledge of one's phenomenal states (as, e.g., Gertler 2012, pp. 102-104 directly concedes[134]) and the latter's – knowledge of one's epistemic states, it is natural to focus on transparency theories when arguing for the luminosity of belief expressed by *KB*.

5.3.1. Byrne's transparency account of self-knowledge

According to transparency accounts of self-knowledge, the agent's knowledge of their beliefs[135] is not acquired by anything close to perception of our "inside world", but rather by attending directly to the content of such beliefs. In the often-quoted words of Gareth Evans, "in making a self-ascription of belief, one's eyes are, so to speak, (…) directed outward—upon the world": to answer the question of whether one believes that *p,* one needs to attend to "precisely the same outward phenomena" as one would when answering the question whether *p* is true (1982, p. 225). Call this

[133] For a more fine-grained taxonomy, see Gertler 2021a.

[134] Russell (1912, pp. 76-80) originally claimed *all* introspective knowledge to be based on acquaintance, but such a strong view is not held by contemporary acquaintance theorists, such as Brie Gertler or Laurence BonJour. For a more in-depth discussion and comparison between acquaintance and transparency accounts of self-knowledge, as well as some historical remarks on the nature of Russell's position, see Tarnowski 2022.

[135] Many authors disagree about the exact scope of the transparency procedure – whether it applies only to beliefs (as e.g., Moran 2001 seems to suggest), or to other mental states as well. Byrne (2018) maintains that *most* or even all of our introspective knowledge is generated by a transparency-like procedure: I shall restrict myself here to considerations of belief and knowledge.

particular process of figuring out what we believe described by Evans *the transparency procedure*[136].

Can utilizing the transparency procedure for one's beliefs yield knowledge? And, if it can, what *kind* of knowledge is generated by following this procedure? There are many responses to these two questions present in the literature. Since the purpose of this chapter is not to adjudicate between them[137], I shall follow one of the most detailed and popular transparency accounts of introspection proposed by Alex Byrne (2005, 2018). As we shall see, Byrne's account is especially promising for our project of finding the theory fulfilling the starting desiderata due to its simplicity, wide scope, and straightforward compatibility with reliabilist and naturalist epistemology.

In short, Byrne's approach to the problem could be summed up as describing the transparency procedure straight-up as a form of *inference*, thereby grounding our capacity for self-knowledge in a naturalistic competence non-controversially possessed and mastered by human agents. More specifically, such transparent inference is described as utilizing a specific set of *epistemic rules*, i.e. rules explicitly guiding one's acquisition of belief expressed by a conditional of the form (Byrne 2018, p. 101):

R: If conditions *C* obtain, believe that *p*.

where the antecedent of R specifies the relevant conditions *C* that need to obtain in order for the one to rationally form the belief that *p* (as stated in the consequent of R).

---

[136] Earlier remarks concerning the transparency of perceptual processes that mirror Evans' description of transparency of belief self-ascription can be found in Moore 1903, p. 446; Ryle 1949, p. 152; cf. also 5.633 and 5.6331 of Wittgenstein's *Tractatus*.

[137] Burge (1996), Moran (2001), and Boyle (2011) are other detailed accounts.

Inspired by an earlier proposal of Gallois (1996), Byrne proposes to describe the agent acquiring transparent self-knowledge of beliefs as an agent trying to follow a rule of the form:

BEL: If *p*, believe that you believe that *p*.

Obtaining higher-order beliefs with the use of BEL is described as follows. On Byrne's account (2018, pp. 101-102), one *follows a rule* if one *knows* the premises of the rule and *tries to follow a rule* if one merely (and perhaps erroneously) *believes* them. When one is in a position to apply BEL, one therefore either believes or knows that *p*, and hence (since we proceed under the assumption of the doxastic nature of knowledge) one's higher-order belief resulting from applying BEL – that *one believes that p* – will come out true regardless of whether one actually follows or merely tries to follow BEL. By this process, one is guaranteed to have accurate higher-order beliefs about their mental states. Moreover, it grounds our introspective process in inferential abilities which are uncontroversially possessed by ordinary agents and the success of applying the rule is properly limited to the first-person perspective. This, in turn, allows Byrne to account for both "peculiar" and "privileged" access to one's beliefs (see Byrne 2018, pp. 108-112), i.e. the intuition that one knows that they have certain beliefs *differently* than others know it of them, and enjoys epistemic *authority* concerning this fact.

Can higher-order beliefs acquired by applying BEL be, however, adequately described as knowledge states? Many of the criticisms pointed against Byrne's account (e.g. Boyle 2011) mention the fact that the inference supposedly performed when one uses BEL is very much unlike any other type of inference that is supposed to yield knowledge. There is neither a causal nor evidential connection between the truth of *p* supposed in the premise and the fact that one believes it, stated in the conditional's conclusion; *p* does not make the fact that *one believes that p* anyhow more likely. Moreover, the explanation of BEL's success explicitly allows one to acquire knowledge

by reasoning through a false step, for *p* might very well be false while the true conclusion is to be granted the status of knowledge. In short, to use Boyle's phrase, "the inference is mad" (2011, p. 230) and does not fit any of the ways standardly associated with justifying an inferential rule.

While Byrne developed detailed responses to such counterarguments (Byrne 2011; 2018, pp. 121-127), their crucial point, which closely aligns his view to externalist intuitions we seek, is emphasizing the unmatched *reliability* of BEL-based inference as a belief-producing mechanism[138]. Note that the BEL rule, unlike other standardly employed rules of inference, secures the truth of the resulting belief irrespective of whether one actually follows or merely tries to follow BEL – in Byrne's terms, this makes BEL *strongly self-verifying*. One may usefully compare BEL with another epistemic rule:

DOORBELL: If the doorbell rings believe that there is someone at the door.

or a classically valid *modus ponens* inference couched in terms of an epistemic rule:

CLOSURE: If *p* and *p implies q*, believe that *q*.

While both of these rules are plausibly valid epistemic rules or, at the very least, good epistemic guides, neither of them produces only true beliefs if applied. For example, unlike BEL, CLOSURE is not strongly self-verifying, for one's false belief in the conditional and its premise may lead one to a false conclusion: in Byrne's terminology it is merely self-verifying, as it yields knowledge only if one follows, not just tries to follow it. DOORBELL is not even that, as one's knowledge that the doorbell

[138] Though Byrne does not endorse a reductive analysis of knowledge in terms of the reliability or safety of beliefs, he does endorse a broadly externalist epistemology that perceives these features as essential properties of beliefs deemed to be knowledge states (Byrne 2018, pp. 103-109).

rings does not guarantee the presence of someone at the door. Though accepting both CLOSURE and DOORBELL might be based on their ability to produce reliably true beliefs, this would be superficial at best, as their reliability as epistemic rules is grounded in the validity of a logical rule of inference or a pattern of abductive, "best-explanation" reasoning. A brain in a vat would be well advised to try to follow CLOSURE and DOORBELL even though beliefs produced by applying them would turn out to be at large false; but this advice would be good only because these rules are fashioned after otherwise justified patterns of inference. On the other hand, a brain in a vat trying to follow BEL would still end up with true higher-order beliefs about their beliefs and could be advised to apply it on these grounds alone It seems that one may require the above traits – "no false lemmas", logical or evidential connection between the premise and the conclusion – from an epistemic rule insofar as the rule is *not* strongly self-verifying and the risk of error arises. If a rule *is* strongly self-verifying, this may be sufficient to justify its use on purely reliabilist grounds. For this reason, also, one need not leave a margin for error for a higher-order belief being the result of applying BEL; the threat to the belief's safety never arises in the first place[139].

5.3.2 Transparency and *KB*

What is the relation between BEL and the *KB* rule? One may plausibly derive a limited, and yet sufficiently strong version of *KB* from the assumption that BEL is a knowledge-generating rule which does not require from an agent anything more than the ability to apply it:

---

[139] In the following section, I shall confront possible objections to this observation. For now, just note that, unlike other cases of inferential knowledge, the possibility of a mistake in applying it is inexistent or at least extremely unlikely.

(**Adjusted *KB***) If *a* believes that *p* and *a* is able to apply the BEL rule to the premise that *p*, then *a* is in a position to know that they believe that *p*.

Is *Adjusted KB* sufficient for the practical purposes the introspection principles are set out to do? In a way, the adjustment (modeled after Das and Salow's [2018, p. 8] modification of *KK*) is proposed to counter the "standard" objections to introspection principles, such as a purported lack of the concept of belief necessary to *form* the higher-order belief at all (e.g., Castañeda 1970) or the lack of computational power necessary to perform relevant deductions: if one is (as I am) skeptical of the power of these arguments for principles concerning epistemic rationality, the provision might just be dropped. On the other hand, if these objections are sound, they are presumably not only pointed against introspection principles, but *all* doxastic-epistemic principles: an agent might not be able to perform apply CLOSURE to the premises *p* and *p entails q* because of their conceptual or computational limitations – the denial of the doxastic closure principle for rational belief does not follow from that. In Levi's (1991) terminology, in our modeling, we should be more concerned with *doxastic commitment* – what agents *should* or *could* come to believe or know if they perform well – rather than *performance,* i.e. their actual belief or knowledge state. If BEL is a good epistemic rule producing only true beliefs that we might be realistically expected to follow, then, just as CLOSURE, it may be said to be a good guide for our rational belief acquisition and change.

On the first pass, Byrne's theory allows us to account for both of our starting desiderata: securing one's strong introspective access to their beliefs as a source of knowledge by appealing to the reliability of the transparency process and on the other hand avoiding the margin-for-error considerations concerning the process' reliability, which allows to describe our second-order knowledge generated by BEL as peculiar and distinct from first-order knowledge gained from experience or inference. Let us now see, how the above discussion translates to the debate surrounding *KK*.

5.4. Transparency of knowledge and *KK*

One thing I demonstrated in the previous section was that Byrne's transparency account provides some plausible grounds for an externalist case for *KB*; what needs to be shown is that this case extends also to *~KK+KB*. I am not the first one to point out the connection between Byrne's theory and the justification for introspection principles – the first such case was made by Das and Salow[140] (2018). Yet, the focus of their approach was to use the transparency account to defend *KK,* not *KB.* As I wish to do the exact opposite, I will now rebut their argument. In the course of this rebuttal, I will also propose some independently plausible augmentations for Byrne's transparency account and demonstrate how it may ground an externalist argument against *KK* different from that of Williamson.

5.4.1. Das and Salow on *KK*

In making their case for *KK*, Das and Salow do not make use of BEL but the epistemic rule KNOW, proposed by Byrne (2012; also 2018, pp. 116-117):

KNOW: If *p*, believe that you know that *p*.

On their account, if we assume that the agent is in a position to apply the KNOW rule and follows it (in Byrne's sense of rule-following), one's resulting belief will be guaranteed to be true in the same way as the belief produced by BEL. Because unlike BEL, KNOW is not *strongly* self-verifying, i.e. the resulting belief comes out true only if the agent actually follows KNOW and not merely tries to follow it, Das and Salow claim that the belief resulting from the application of an epistemic rule qualifies as

[140] A similar argument to Das and Salow's, though not explicitly based on Byrne's account, is made in (McHugh 2010). Byrne himself seems to remain agnostic with respect to whether his account provides any support for introspection principles (2018, p. 116).

knowledge *only if* the agent qualifies as following the rule in Byrne's sense. Hence, whenever one comes to believe that one knows that *p*, although one does not know that *p*, their belief is a result of merely *trying to follow KNOW* which, according to Das and Salow, is insufficient to generate knowledge (2018, p. 10). If one, on the other hand, *follows KNOW*, one is epistemically guaranteed to succeed – and *KK* (in a qualified form analogous to *Adjusted KB*) gets validated.

As we may already see, the case based on KNOW Das and Salow propose differs significantly from the simple reliabilist justification provided above for *KB* based on BEL. Since KNOW produces true beliefs only if followed – it is self-verifying, but not strongly – the question of why it could be said to be knowledge-producing no longer can be answered on the grounds of the exceptional ability to produce true beliefs. A brain in a vat will no longer easily gain higher-order knowledge; it is also quite plausible that a purely statistical assessment of reliability will qualify only some – more fortunate or epistemically cautious – actual world agents as forming reliably true beliefs by applying KNOW.

Das and Salow support their case by likening the knowledge obtained by following Byrne's transparent rules to the knowledge obtained by following rules like CLOSURE[141] (2018, p. 5). Their argument is fairly simple: given that CLOSURE, like KNOW, does not generate knowledge if one does not *know* the premises *p* and *p entails q*, one's epistemic base for gaining knowledge through epistemic rules needs to be *following*, not *trying to follow a rule*. If the agent, for example, merely believes but does not know that *p entails q*, one's belief in *q* based on trying to follow CLOSURE can come out to be either false or unsafe (2018, pp. 11-12). Because we should count all of the rules governing the acquisition of inferential knowledge as having the same type of

[141] Instead of CLOSURE they use OR, epistemic rule based on disjunction introduction. The choice of the rule does not matter here.

epistemic base (by what Das and Salow call *Generality Constraint* [2018, p. 11]), it follows that all epistemic rules should count as knowledge-generating *only if* the agent follows them, not merely tries to; this, in turn, gives them the right to claim that only beliefs resulting from *following* KNOW may count as knowledge, and validate *KK*.

What should we think of this story? One question that should arise in the mind of a careful reader of the previous section is what is, on Das and Salow's account, a good basis for higher-order knowledge of one's beliefs. Treating BEL as knowledge-generating only if the agent in question *knows that p* feels weird: a brain in a vat intuitively *can* obtain knowledge of its beliefs by applying BEL even if its first-order beliefs are always false. By Das and Salow's argument from analogy this would be however the case; if we treat CLOSURE analogously to KNOW as rules producing inferential knowledge that need to have a uniform type of base, then BEL should have the same one as well. In line with what I've said previously, I suggest that it's better to distinguish various kinds of inferential knowledge as being formed on different epistemic bases, as the argument for the adoption of a specific epistemic rule varies from brute reliability (BEL) to being a form of valid logical inference (CLOSURE). Inferential knowledge need not have one unified type of base in Das and Salow's sense, as "being inferential" itself seems to be a genetically individuated species of knowledge: rules of knowledge-producing inference can vary in their quality, reliability, and plausibly ultimate justification. Moreover, underlining the peculiarity of BEL as having a distinct type of base (i.e. trying to follow it) sufficient for reliably producing true beliefs crucially allows for making higher-order knowledge of one's own beliefs special and immune to *MAR*-related worries, even if we are, along with Williamson and many other epistemic externalists, compelled to accept *K-MAR* for first-order knowledge.

5.4.2. KNOW and first-person knowledge-belief collapse

How then are we to think about our self-ascriptions of knowledge? Even if we discard the analogy-based justification for *KK* provided by Das and Salow, upholding KNOW would remain problematic if we are, at the same time, willing to grant the success of Williamson's weaker argument against *KK* based on the knowledge of *MAR*. Accepting both BEL and KNOW as tools in our cognitive repertoire leads to the following trouble: one's self-ascriptions of belief based on trying to follow BEL need to align with self-ascriptions of knowledge based on trying to follow KNOW, as both rules use the same premise. In effect, this results in the *knowledge-belief collapse* from the first-person perspective – one cannot believe that one believes that p without simultaneously believing that one knows that p. If the agent is in a position to apply the KNOW rule, then whenever they try to follow it and believe that p, regardless of whether they know p, they end up with a higher-order belief that they know that p. Therefore, accepting KNOW as an epistemic rule justifies the introduction of another introspection principle, *BK*[142]:

(*BK*) $Bp \rightarrow BKp$ (*If one believes that p, then one believes that one knows that p).*

Accepting *BK* violates, however, our initial assumption that *MAR* applies to first-order knowledge and that the agent may know this fact about their knowledge: in essence, it invites Williamson's weaker anti-*KK* argument within the scope of the belief operator. If we assume that Mr. Magoo believes that it is not 6 o'clock in α_0, he may, by *BK* and knowledge of *MAR* for first-order knowledge, come to believe that it is not 6 o'clock in α_1 and repeat the same process up to coming to believe that it is not 6

[142] This principle is introduced by Stalnaker (2006) as "strong belief", and is also accepted in a qualified form by Greco (2014). On the other hand, Bird and Pettigrew (2021) christen it with a derogatory (yet, in my opinion, apt) label "arrogance".

o'clock in α_{120}[143]. More concisely, accepting KNOW and *K-MAR* for first-order knowledge leads one to accept the following soritical epistemic rule:

SOR: If *p* is true in α_i, believe that *p* is true in α_{i+1}.

For if one believes that one knows that *p* is true in α_i, by one's knowledge of *MAR* for first-order knowledge one can deduce that *p* is true in α_{i+1}, which in turn becomes (by the assumption that one may always warrantably try to follow KNOW when one believes its premise) the basis of the belief that one knows that that *p* is true in α_{i+1}. This leads the agent to believe that *p* is true at any given case α in an arbitrarily long sequence α_i-α_n, which is irrational.

To recapitulate, even if we are to disagree with Das and Salow in that there is a close analogy between justifying *KK* with KNOW and *KB* with BEL, KNOW itself remains problematic, as it allows justifying *BK* on analogous grounds, which, in turn, makes the self-ascribed beliefs susceptible to Williamson's weaker anti-*KK* argument. Should the above reasoning lead us to give up hope in using Byrne's framework to justify introspection principles or to modify it by abandoning or modifying KNOW? In the next section, I shall argue that it rather should prompt us to modify Byrne's theory, which will allow us to draw a principled distinction between *KK* and *KB* needed for our project.

5.5. Divorcing knowledge and belief

The last section presented an argument to the effect that accepting KNOW should lead us to reject strong introspection principles on Williamsonian grounds: if KNOW is an essential element of Byrne's framework, I failed in my project. But is the status of

[143] An argument to the same effect, aiming at extending Williamson's anti-*KK* argument to other introspection principles, is used by Hawthorne and Magidor (2009).

KNOW as central or as secure as BEL's? Shouldn't we choose, from an externalist point of view, a more cautious epistemic rule guiding our self-ascriptions of knowledge?

Byrne himself makes his case for KNOW rather brief, by pointing out an analogy between Evans' transparency procedure for self-ascriptions of belief and self-ascriptions of knowledge:

"If someone asks me 'Do you know that it's raining?' I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Is it raining?'. This sounds equally plausible as the original claim about belief, suggesting that I know that I know that it's raining by following [KNOW]" (Byrne 2018, p. 116).

There is a good reason to distrust the force of this specific epistemic-conversational pattern. Why? Because there is an important point of disanalogy between "Do you believe that *p*?" and "Do you know that *p*?" questions: while the first does not involve any non-trivial pragmatic assumptions about the truth of *p*, the second one, in normal circumstances, involves the speaker presupposing that *p* is true[144] (i.e., is knowledge-apt in the first place). If we employ a transparency procedure when answering such questions, this fact could be easily explained away by this pragmatic assumption: what one queries in asking such a question is our state of mind – whether we believe that *p* – not our epistemic warrant or reliability – the speaker is pragmatically expected to know that *p* already. On the other hand, the question "*Do you only believe that, or do you know it?*" is *not* a nonsensical one, and the agent may formulate a sensible answer to it. If I (cautiously) assert that *p*, my interlocutor may direct this question to me. In answering them, I should attend to phenomena peculiar to the epistemic status of my belief, such as the reliability of my source, instead of

[144] Of course, if it is not used to *challenge* the hearer's previous assertion that *p* in a way discussed in Chapters 2. and 3.

reiterating the transparency procedure; the emphasis is now on my reliability, not on my opinion.

The troubling consequences of embracing KNOW discussed above mostly have to do with what I dubbed the "first-person collapse" of knowledge and belief. While many philosophers tend to like the idea of such collapse, I think it results, in a way, from a mistaken phenomenological intuition. When defending a similar thesis, it is sometimes invoked that, from a first-person perspective, believing "feels just like knowing" (Greco 2015b, p. 180; see also Stalnaker 2006, pp. 179-180; Tokarz 1993) or is "an attitude (...) which is, for all one knows, knowing" (Williamson 2000, pp. 46)[145]. While we might concede that in one cannot properly internally distinguish between merely believing and knowing that *p*, it does not follow that one thereby always takes one's beliefs to be knowledge states, just that one usually cannot properly determine whether or not a certain belief of theirs constitutes knowledge. In other words: phenomenal indistinguishability of actual cases of knowledge and mere belief proves just that one can be wrong in thinking that one of their beliefs is or is not knowledge, not that one does not make any (perhaps erroneous) internal distinctions between the two states.

Reasons to think that we not only can but also *do* distinguish between our own beliefs as mere instances of belief and as knowledge states have been already brought up in this dissertation in the form of arguments against the knowledge norm of belief, sometimes formulated in a wording close to that of "knowledge-belief collapse", e.g. in saying that "if one believes that *p* one is thereby *rationally committed* to taking one's belief to be knowledge" (Huemer 2007, p. 145). In the previous chapter, we found that

[145] Williamson does not endorse this specific view about belief in *Knowledge and Its Limits* (2000, pp. 46-47; for a full endorsement of similar view, see Stalnaker 2006 and Lenzen 1979); instead, he usually characterizes a belief that *p* to be a "disposition to act as if one *knew p*" (see, e.g., his 2017).

accepting this norm is neither sufficient nor necessary to demonstrate the irrationality of Moorean beliefs (which is sometimes cited as its justification). More directly, I noted already in Chapter 2. that accepting this norm together with any sort of doxastic characterization of assertion makes seemingly felicitous assertions of the form:

(BWK) I believe that *p*, but I don't know that *p*.

irrational by making them essentially equivalent to Moorean assertions[146]. Similarly, denying that one needs to treat one's beliefs as one's knowledge was implicit in the argument for context-sensitivity of the norm of assertion, as I argued that certain conjunctions of the EOMP form may be felicitously asserted and rationally believed.

On the ground of pure epistemology, denying the collapse squares well with the externalist position I adopted here. From the reliabilist standpoint, it is perfectly coherent to hold that we may permissibly believe something, while at the same time thinking that this belief may not constitute knowledge, because we are unsure whether it is reliably formed. Think again of Mr. Magoo observing the strange clock. When he comes to believe at 4.50 that it is not 5 o'clock, he may remain agnostic as to whether his eyesight is good enough to grant him the knowledge that it is not 5 o'clock, and yet fully believe that the world is the way it presents itself to him, even if only for the fact that there is no better data he could rely on in his practical reasoning. If his skepticism

[146] One may, as Greco 2014 does for example, maintain that in such cases one self-ascribes themselves a state of *weak belief* (understood along the lines of Hawthorne et al. 2016), while the first-person collapse holds for *strong,* or *proper belief*. I don't find this move persuasive, for even if we grant the existence of weak beliefs as Hawthorne et al. understand them, they are essentially weaker, and usually self-ascribed with "I think..." or "I guess..." clauses (Dorst, Mandelkern 2022), than the doxastic states in question. Of course, one cannot forbid another to simply define "strong beliefs" as those for which the collapse thesis holds, but I doubt that in this way we learn anything useful or general about all belief states.

towards his perceptual abilities is unwarranted, his first-order belief constitutes knowledge (but his *second-order* belief is wrong), if it is warranted – then it does not (and second-order belief is right). But while he has no means to assess which of the scenarios he is in, and hence cannot truly distinguish "from within" whether he merely believes or knows that it is not 5 o'clock, it does not seem like a conceptual truth or a form of irrationality on his part to stick to believing what seems to him true.

What should we then do with KNOW if we reject the intuition of the first-person knowledge-belief collapse? I suggest that staying true to externalist intuitions, we may propose the following Byrne-style epistemic rule to guide our self-ascriptions of knowledge:

KNOW*: If you believe that p and your belief that p has high epistemic credentials, believe that you know that p.

KNOW* is a transformed CONFIDENCE rule, proposed by Byrne (2018, pp. 119-121) to guide our self-ascriptions of high (and low) confidence in belief. By "high epistemic credentials" ascribed to the belief I mean, like Byrne, relevant evidence about the source of the belief and reliability of the belief-forming mechanism. Importantly, unlike in KNOW, self-ascriptions of knowledge are here "broadly Rylean", i.e. based on the assessment of the reliability of our cognitive faculties, perception, memory, etc. that is in principle accessible also to others. This feature allows us to accommodate the Williamsonian intuition that the imprecision built into our belief-forming mechanisms may prevent us from knowing that we know something even if we do know it. Mr. Magoo may very well come to believe that it is not 5 o'clock by looking at the clock showing half past 4 from a distance, but still, after reflecting on his poor sight and discrimination abilities, refrains from believing that he knows that, as his epistemic credentials do not meet the desired threshold.

All this seems quite natural for an epistemic externalist. If an externalist takes themselves to know that knowledge that *p* requires sufficient margin-for-error (as *per* Williamson's second anti-*KK* argument), they may be expected to be more conservative in their self-ascriptions of knowledge than that of belief. According to them, knowing that one knows, unlike knowing that one believes, requires also knowing that their first-order belief meets relevant externalist criteria[147], which feature here under the label of "high epistemic credentials". Only when one believes or knows that their belief satisfies all these criteria, they may conclude by KNOW* that they know that *p*. Unlike BEL and KNOW, KNOW* is not even self-verifying – the protagonist of some modified Gettier scenario could arrive at a false conclusion that they know that *p* despite *following* KNOW*. However, in opposition to these rules and much more like DOORBELL, KNOW* is abductively justified: that one believes that *p* with high epistemic credentials is just good evidence for that one knows that *p*. Quite easily we can then classify KNOW* as knowledge-generating, without the troubling consequence of knowledge-belief collapse, upholding skeptical conclusion that we are never in a position to know that we know anything or denying *KB*.

5.6. Conclusion

The importance of the conclusions of this chapter is perhaps best commenced by characterizing the audience the chapter is aimed at – a portrait of a sympathetic reader.

---

[147] One may protest that this reasoning commits an intensional fallacy, for according to externalist standards it does not matter whether the agent's belief that they know that *p* is supported by such robust evidence, but only that it is properly formed (Okasha 2013). However, if an agent *is* an externalist and *believes* that first-order knowledge requires e.g. Williamsonian margin for error, then one cannot believe that they know that *p* without believing that their first-order belief meets MAR on pain of doxastic inconsistency (see Bird and Pettigrew 2021, pp. 1726-1727). Hence, since to run Williamson's argument for first-order knowledge we *need* to assume that the agent *knows* or *believes K-MAR*, we may stipulate that such conditions obtain.

I take it that such a reader, finding themselves with externalist epistemic intuitions, has both sympathy towards Williamson's epistemic project and *MAR*-style analysis of knowledge, and finds *KK* implausible to begin with. The reader may then feel the temptation to spill this skepticism over all introspection principles. When this skepticism takes hold, we find ourselves in serious trouble when it comes to the content of this dissertation, trying to account for the irrationality of Moore-paradoxical beliefs (as explained in Chapter 4.) and explain rational coordination between speakers, which seems to require at least common knowledge of conversationists' beliefs (as explained in section 2. and employed in Chapter 3.). In the chapter, I argued that this temptation can be resisted, and important results saved, on principled grounds without steering too far away from Williamson's externalist framework.

In section 3. I provided a natural way of restricting Williamson's *MAR* principle to first-order knowledge and showed how this restriction allows for both rejecting *KK* and maintaining *KB*, a combination of views labeled *~KK+KB* thesis. I then proceeded to argue (in section 4.) that Alex Byrne's transparency framework provides sound philosophical grounds for *~KK+KB*. *Contra* Das and Salow (2018), I argued (in section 5.) that an independently plausible externalist modification of Byrne's theory does a good job of explaining why *KK* might fail, but *KB* remains true. Of course, one might challenge the choice of Byrne's framework as arbitrary, given the sheer number of views about the nature of self-knowledge present on the market. As I explained, however, this view is quite natural for the epistemic externalist, as it grounds epistemic success of self-knowledge in the reliability of transparent inferences which, in turn, remain quite unmysterious and pair well with naturalistic views of the human mind.

I hope that by the end of this chapter, the externalist reader would then be content to agree that, while "[t]o know that you know something is to perform a (...) great epistemological feat, which is not comparable to just knowing it" (Kripke 2011, p. 34), knowing that you believe something requires no such hard work. This, in turn, allows us to save the important practical consequences of introspection principles by

defending *KB* – and, per the results presented in Chapter 4., justify the claim that Moorean beliefs are irrational.

Closing Remarks and Open Questions

In this last chapter, I will briefly recapitulate the main argument of the dissertation, as it runs throughout chapters 1-5, and offer some closing remarks on its content. I will also mention five open problems that were side-lined in them and suggest how they might be resolved in future work.

6.1. The main argument of the dissertation

The present monograph was constructed as an in-depth attempt at providing a uniform answer to Moore's Paradox, encompassing its occurrence both in speech and in thought. As its five chapters deal with many intricate subjects, some of which were less directly connected with its guiding topic, it is worth adopting a bird's-eye view of them to see clearly the exact contribution of each of them to our central issue.

In Chapter 1., Moore's Paradox was defined as the following falsidical paradox:

MOORE'S PARADOX:

(a) In typical circumstances, *p* may be rationally believed and felicitously asserted if it is true.

(b) Moorean sentences (e.g., OMP and CMP) may be true.

(c) [from (a)-(c)] In typical circumstances, if Moorean sentences are true, they may be rationally believed and felicitously asserted.

(d) In typical circumstances, Moorean sentences cannot be rationally believed and felicitously asserted, even if they are true.

where Moorean sentences were characterized (though not *defined*) as consistent sentences meeting the below tests:

(THIRD) It is typically irrational or infelicitous to assert/believe/… *s* in the first-person present tense, but not in the third-person present tense.

(PAST) It is typically irrational or infelicitous to assert/believe/… *s* in the first-person present tense, but not in the first-person past tense.

(SUPPOSE) It is typically irrational or infelicitous to assert/believe/… *s* in the first-person present tense, but not to suppose it.

As noted, the plausible source of this puzzle, given the accuracy of Moore's observation expressed by (d), is the intuitive principle expressed by (a). The sound method of resolving Moore's Paradox is therefore to propose an alternative constraint on rational belief and felicitous assertion than the truth of its object sentence, provide its plausible philosophical justification, and demonstrate, how it allows to show irrationality of believing or asserting Moorean sentences. The two central questions that need to be answered to solve Moore's Paradox thus stated are as follows:

(Q1) What condition *C* constrains felicitous assertion of some sentence *p* and is such that Moorean sentence is typically not *C*?

(Q2) What condition *C′* constrains rational belief in some sentence *p* and is such that Moorean sentence is typically not *C′*?

Based on the diagnosis of certain problems of two historical solutions – Moore's own and Wittgenstein's – I stated my thesis that these two questions should be given a uniform answer. Following Shoemaker (1995), I put the central thesis of this dissertation as follows:

(**Priority of Belief**) The infelicity of Moore-paradoxical assertions is explained by the irrationality of Moore-paradoxical beliefs.

and hypothesized that the plausible link between belief and assertion needed for such an explanation to work takes the form of the following "bridging principle":

(**Bridging Principle**) One can warrantably assert only what one can rationally believe.

given that assertions that are obviously normatively defunct and unwarranted are judged to be infelicitous. In effect, the condition *C* in question (Q1) was hypothesized to be that of *rational belief*, which in turn was to be constrained by *C'* as mentioned in (Q2).

Chapter 2. explored this hypothesis in more depth from the perspective of contemporary discussions concerning norms of assertion. Firstly, I considered alternative accounts of assertion that do not align with the Bridging Principle, such as those proposed by Weiner (2005), Douven (2006), and Lackey (2007), and those suggested by Hindricks (2007) and Bach (2008). I argued that these accounts, apart from their independent problems, cannot properly account for the infelicity of

Moorean assertions. This discussion left out three views: Williamson's (2000), Kvanvig's (2009), and Stanley's (2008) as possible alternatives. These proposals propose, in turn, *knowledge*, *justified belief,* or *epistemic certainty* as a placeholder for *C* in an answer to (Q1); as all of these conditions entail rational belief, they align with my hypothesis. I evaluated these proposals with respect to their handling of a variety of Moorean felicity data, noting that the infelicity of certain variants of Moorean paradoxical constructions, such as EOMP or COMP, is discourse-sensitive, i.e. varies with respect to its topic. In response to this, I proposed a Flexible Assertion Schema:

(Assertion Schema) S may assert *p* in discourse *D* only if S's belief that *p* expressed by such assertion meets justification standards relevant for *D*.

which, when the standards relevant to the given discourse are specified, allows one to derive a norm of assertion for a given discourse. I showed, how such schema is supported by consideration of epistemic expectations of speakers concerning a given topic (following Goldberg 2015), and proposed a few specific cases of how a concrete norm might be derived from it, e.g. for future-directed and aesthetic discourse. Given that the lower bound of admissibility allowed by the schema is that of rational belief, it allows for an explanation of Moore-paradoxicality at the level of assertion consistent with Priority, as well as accommodates the strengths of all mentioned views.

Though Chapter 3. may be regarded as a detour from the central investigation, it aims at answering a very important question of whether Moore-paradoxicality is only exhibited by assertoric speech acts – and if not, whether it threatens the Priority-based explanation. Given that both the Priority thesis as well as the presentation of Moore's Paradox itself presented in the first chapter characterize it essentially as a problem of *assertion* and *belief*, the presence of supposedly Moorean non-assertoric speech acts

may be a problem for the uniformity and generalizability of the proposed explanation. In this chapter, I followed the principles of Bach and Harnish's descriptive taxonomy of speech acts by the type of attitude they conventionally express as well as Stalnaker's framework for describing the type of effect on the conversation these speech acts have in terms of their interaction with the common ground. This allowed me to substantiate the assumption made in Chapter 2., concerning the grounding of normative properties of assertion in the possible fulfillment of expectations the audience has towards the assertor, as well as characterize such expectations towards the speaker who utters any type of speech act. The central hypothesis of the chapter described the universal effects a speech act has on the common ground as follows:

(Central Hypothesis) By performing the speech act α, conventionally expressing mental state M with a satisfaction condition s, the speaker S proposes to add to the common ground:

(B) the proposition that s occurs or will occur.

(R) the proposition that *S is in M*.

This hypothesis was shown both to explain the infelicity of examples of supposedly Moorean non-assertoric speech acts provided in the literature and to be independently grounded in theoretical reflection present in the linguistic and philosophical literature on the pragmatics of commissive and directive speech acts. I also demonstrated, how this hypothesis can be adjusted to explain related infelicity of examples involving expressives and declarations, as well as cases of conversational pretense.

The impact of the content of Chapter 3. on the overall discussion of Moore's Paradox, as noted there, depends, in my opinion, largely on one's approach toward

the semantic-pragmatic input of explicit performative verbs. Given that the tests for Moore-paradoxicality require uniformity concerning which speech act is being performed with the use of sentences in first-person present tense and its third-person and past-tense versions and to allow supposition, I said that it makes sense to follow the idea of Bach and Harnish and think of explicit performatives of the form "I α that..." as cases of α*'s* performed *indirectly* by the means of a self-verifying assertion. If that were the case, some sentences containing explicit performative verbs would plausibly count as cases of Moorean assertion and their infelicity would be given an analysis in terms of the assertion's warrantability analogous to that of standard Moorean assertions, given the truth of the Central Hypothesis. On the other hand, pairs of speech acts performed *via* interrogative or imperative sentences and the assertions that directly or indirectly deny (B) or (R) condition of the Central Hypothesis would still be classified as infelicitous, though not Moorean. If one, however, prefers an Austinian view, according to which explicit performatives are just that – *explicit* performatives – I suggested that one may treat Central Hypothesis as an explanation of infelicity of the problematic constructions discussed in the literature regardless of whether they should or should not classify as properly "Moorean".

Part I of the dissertation, composed of chapters 2. and 3., as a whole can be seen as providing the argument that the right answer to (Q1) points towards rational belief as a restriction on what can be warrantably asserted, and that this fact may account for the infelicity of commonly discussed examples of Moorean speech acts. Part II dealt with supplementing this picture and answering (Q2) by explaining the irrationality of Moorean thought, which, if the analysis provided in Part I was right, should be seen as central to solving Moore's Paradox as a whole.

Chapter 4. brought two considerations concerning Moorean irrationality. Firstly, I briefly motivated the centrality of belief in Moorean thought, that is – why belief should be considered the only "Moorean" mental state. Secondly, I compared two ways in which different specific explanations of Moorean irrationality proceed: the

Introspectionist strategy, which takes Moorean beliefs to violate the rational constraints on self-knowledge, and the Self-defeat strategy, which argues that Moorean beliefs are irrational because they are false when believed. An often-repeated argument for the latter strategy (employed by defendants of the view that belief should aim at either truth or knowledge) is that it allows one to demonstrate the irrationality of Moorean beliefs without committing oneself to implausibly strong principles of rational belief, such as the controversial *BB* principle employed by introspectionists:

(*BB*) $Bp \rightarrow BBp$ (*If one believes that p then one believes that one believes that p*)

I analyzed this argument as resting on the intuitive desideratum guiding the theoretical preference between different solutions to Moore's Paradox in belief:

(Innocence) The preferred solution to Moore's Paradox needs to make only *minimal assumptions* concerning the characteristics of rational belief.

To precisely assess whether the theoretical commitments of Self-defeat theorists are more "innocent" than those of the introspectionist, I used the tools of epistemic and doxastic logic to approximate the minimal assumptions about rational belief these two approaches are committed to. Following the result of Adam Rieger (2015), I have taken the minimal logic of self-defeater's commitments to be characterized by the doxastic logic *K5c,* that is a normal doxastic logic *K* supplemented with the following axiom:

(*5c*) $B{\sim}Bp \rightarrow {\sim}Bp$

I also argued that an analogous system of bimodal logic (labeled *B~K*) can be said to express the commitments of the knowledge-norm theorist (who also takes self-

defeat as their preferred way of demonstrating the irrationality of Moorean beliefs), which replaces *5c* with the following axiom:

(*B~K*) $B{\sim}Kp \rightarrow {\sim}Bp$

The minimal system of doxastic logic corresponding to introspectionist commitments was, as standardly assumed, taken to be the system *KD4* – logic *K* with *BB* and *D* (the principle expressing the idea that one cannot rationally hold contradictory beliefs) added as axioms.

While *K5c* is a weaker system than *KD4* and therefore the Innocence-based argument for Self-Defeat is sound, it is so only insofar as OMP and CMP are treated as the *only* Moorean sentences such strategies need to account for. In the rest of the chapter, I introduced two problematic sentences, which (as I argued) meet a theory-neutral characterization of Moore-paradoxicality provided in Chapter 1.: the *anti-expertise* and *iterated Moorean sentences*. Both of these sentence types proved to be problematic for Self-Defeat, as they *can* be rationally believed by agents satisfying the constraints of *K5c* and *B~K* logics (and, moreover, they are not even self-defeating in such logics), while the same is not true for *KD4,* "introspective" agents. The argument from Innocence, therefore, turns out to be double-edged: "weak" theoretical commitments of self-defeat strategy, while initially desirable, prevent it from accounting for all cases of Moore-paradoxicality. Hence, I concluded, that a solution along the Introspectionist lines – crucially, one which insists on *BB* as a principle true for rational belief – is to be preferred.

While Chapter 5. may seem, at first, least concerned with an analysis of Moore-paradoxicality, its main aim in a wider context is to choose a theory that may provide sound philosophical justification for introspection principles for belief, such as *BB*, which allows us to solve Moore's Paradox along the lines defended in Chapter 4. To do this, I argued, we need a compelling account of introspective knowledge that

escapes the arguments directed against introspective principles and provides a natural grounding for their truth.

As I noted, while it is natural to defend introspection principles from the epistemic internalist standpoint, they are usually contested on the externalist side of the debate, most importantly due to the popularity of anti-luminosity arguments formulated by Timothy Williamson (1996, 2000) based on the "margin-for-error" principles. I decided to demonstrate that one needs not to abandon the externalist intuitions to defend at least some of the introspection principles against Williamson's attack, most importantly – the principle *KB*:

(*KB*) $Bp \rightarrow KBp$ (*If one believes that p then one knows that one believes that p*)

which entails *BB* and hence allows to demonstrate the irrationality of Moorean beliefs. The defense proceeded in two steps: firstly, I have shown that one may defend *KB* by restricting the scope of Williamson's margin-for-error principle only to first-order knowledge states, in line with the observation that his justification for *inexactness* of knowledge, while plausible for perceptual knowledge, does not intuitively extend to knowledge gained through introspection. This limitation, crucially, still allows one to disprove the *KK* principle, denied by epistemic externalists, if we admit that such restricted margin-for-error principle is known by the agent. Secondly, I set out to find an account of introspective knowledge that aligns with broadly externalist views and allows one to both uphold the *KB* and reject the *KK* principle. Such an account, I argued, can be provided within Alex Byrne's (2005, 2018) *transparency* framework, which treats self-knowledge as knowledge obtained through *epistemic rules*, most crucially the following BEL rule:

BEL: If *p*, believe that you believe that *p*.

As Byrne's justification for the claim that BEL is a knowledge-generating rule appeals to the reliability of BEL's application, and, as I demonstrated, it can be used to justify the *KB* principle, I took it to meet the desired characteristics. Furthermore, I extended and modified Byrne's position to answer Williamson's challenge and argued, *contra* Das and Salow (2018), that Byrne's theory does not justify the *KK* principle and hence is consistent with a more moderate interpretation of Williamson's argument. This modification crucially involved making self-ascriptions of knowledge more "cautious" and not perfectly aligned with self-ascriptions of belief, i.e. allowing one to self-ascribe belief that *p* without at the same time self-ascribing knowledge that *p*.

If I am right in my analysis, these five chapters taken together demonstrate, that the right solution to Moore's Paradox can be given in fairly straightforward terms by combining the sufficiently strong, doxastic norm of assertion (as defended in Chapter 2. and extended in Chapter 3.) and introducing introspective constraints on rational belief (as defended in Chapter 4. and extended in Chapter 5.). While the intricacies of respective subjects led me to side with or develop more concrete versions of these two claims that I find most plausible – the *Flexible Assertion Schema*, the *Central Hypothesis* concerning effects of speech acts on the common ground, a specific interpretation of Byrne's transparency account of self-knowledge – I believe that I have shown that at least in its broad strokes this approach to Moore's Paradox has its indubitable strength.

6.2. Some open questions

While this dissertation covered a lot of material and, plausibly, most of the attempts at solving Moore's Paradox offered in the literature, nevertheless certain questions related to it were put aside. Here, I want to briefly list five of them (going in

a topic-progression intuitively related to the themes of subsequent dissertation chapters), to both offer the reader a guide to other topics connected with Moore-paradoxicality that they will not find here and, if possible, to hint at an answer that might be given along the lines of this thesis' methodological assumptions.

- **Cases of "atypical", felicitous Moorean assertions and rational Moorean beliefs**

Problem: As briefly noted in Chapter 1., there are cases in which one seemingly *can* felicitously assert or rationally believe a Moorean sentence. Turri (2010) presents a case of the eliminative materialist, who holds that there are no beliefs and felicitously asserts "*p*, but I don't believe that *p*, since there are no beliefs". Borgoni (2015), Fileva and Brakel (2019), and Gertler (2021b) discuss *implicit bias* and *akratic* cases in which one self-ascribes a belief based on the pattern of their own behavior which does not match their conscious judgments, leading them to believe a Moorean conjunction; Pruss (2012) offers a similar case, where this belief self-ascription is adopted based on the testimony of a psychoanalyst. Other structurally similar cases were reproduced, e.g., for the protagonist holding differing views on the metaphysical domains of ordinary objects and the conceptual structure of belief (Frances 2016), doxastic voluntarism (Kvanvig 2009) or dialetheism (Williams 2015; for a survey of some more, see Hajek 2007).

In Chapter 1. I noted that these examples are "atypical" enough not the threaten the general point made by Moore's observation: it suffices for standard cases of Moorean assertions and beliefs to be "typically" infelicitous or irrational irrespective of their truth to merit serious philosophical puzzlement. These examples, however, seem to have something in common, and they all do not fit the universal explanation: what is a universal source of the fact that we are not particularly baffled by such cases, while we are when it comes to "standard" Moorean assertions and belief, is an interesting puzzle in itself.

Possible answer: From the perspective of this dissertation explanatory strategy, the interesting thing that unites all the mentioned cases is the fact that in all of them the protagonist either publicly challenges one of the postulated norms, or is otherwise saliently exempt from them. The eliminativist refuses to accept the doxastic norm of assertion and any norm governing rational belief; the dialetheist – that believing contradictions is irrational; Borgoni's and Gertler's cases involve a protagonist who openly admits that they hold an irrational belief that they are unable to change. The suggestion is that such examples crucially involve salient breakdown of one or the other principle that we typically employ in assessing the rationality of one's belief or warrant of one's assertion; this, perhaps, suffices for them to be excused from breaking such rules, as the reason for why they do so is salient, though their assertions and beliefs are still defective (see Williamson 2013 pp. 341-342 for a suggestion in this direction). Still, plausibly modeling conversations in which these rules are knowingly broken requires additional work.

- **Demonstrative and indexical reference and Moore-paradoxical utterances**

Problem: Another interesting set of cases of supposedly felicitous Moorean assertions, which (though related in spirit to those mentioned above) as of yet were not explicitly discussed in the literature, involves the indexical nature of Moorean sentences. Consider a case, in which a humorous philosopher of language puts the following statement at the beginning of his will:

(1) I am dead, but obviously I don't believe it now.

expecting it to be read after his death. When the time comes and the will is read after his funeral, (1) seems true and perfectly felicitous, even though it fits the standard form of an omissive Moorean sentence. Analogously, one may imagine a colleague who, knowing that I have mistaken the time of the department meeting but is unable to

reach me, decides to write the following sentence on a note and hang it on my office door:

(2) The department meeting is today at 6 p.m., but I believe it's not.

again producing a Moorean sentence that does not sound infelicitous and can be perfectly informative.

While the problem runs deeper and concerns, in general, our understanding of reference-fixing for indexical expressions (as more elaborate versions of so-called "answering machine cases"), the fact that similar cases can be reproduced with Moorean sentences calls into question my syntactic characterization of the problem and invites theoretical reflection on their context-dependent nature.

Possible answer: As indicated above, the issue concerning similar utterances is well established in the literature concerning indexical reference. In David Kaplan's popular framework (1989), the character (*linguistic meaning*) of "pure" indexicals such as "I", "now" or "here" always picks out the agent, location, and time of the context of the utterance, which, in turn, is thought by Kaplan to simply be the context of utterance's production. Yet, as both examples above demonstrate, this approach seems to go against our semantic intuition in at least some cases – for both utterances seem to be true and felicitous. Clearly, in (1) "I" seems to refer to the deceased philosopher and "now" to the time of the reading of the will, though at this time the philosopher is dead; in (2), the intuitive referent of "I" is still me, though the note was produced (without my knowledge) by my colleague.

To accommodate similar cases, a variety of different alternatives to standard Kaplanian treatment were proposed. One such compelling idea is suggested by Stefano Predelli (1998), who argues that instead of assuming that the context of utterance is automatically fixed by the circumstances of its production, we should instead take the "context of intended interpretation" to be semantically binding. In this reading, both (1) and (2) get intuitive readings and can be regarded as felicitous as

essentially expressing (at the moment of production) a future tensed or third-personal belief (as the context of the intended interpretation of (1) is timed after philosopher's death, and (2) takes me to be its agent). This proposal, however, comes with important caveats. Apart from the general problems with intentionalism about indexical reference (for discussion see e.g. Corazza et al. 2002) accepting Predelli's solution would mean imposing further restrictions on the syntactic characterization of Moorean sentences, or forcing the characterization of "being Moorean" in propositional terms (taking into account various caveats mentioned in 1.5.3.). To avoid this issue, one might also consider the hypothesis that such utterances contain "deferred", demonstrative uses of "now" and "I" (cf. Ciecierski's 2022 approach to the answering machine paradox) and posit that such uses should be represented differently at the level of syntax, not only semantics, which would ultimately leave the syntactic characterization of Moore-paradoxicality unscathed.

- **Discourse-specificity of norms of assertion**

Problem: Chapter 2. of this dissertation ends with a reflection on the limitations of the proposed form of Flexible Assertion Schema; I indicated, that a fully fleshed-out theory of assertion following the route established in the chapter should provide a precise taxonomy of different discourses, which are connected with differing epistemic expectations and normative demands of conversation participants. One might question whether such theory is, actually, feasible; given that the term "discourse" is loosely applied across nearly all social sciences and humanities, the specter of *ad hoc*-ness may start to loom. For example, one might wonder whether a conjunctive assertion (like typical Moorean ones) should be held to epistemic standards specific to the discourse of one of the conjuncts, or should both conjuncts be evaluated separately? Or is the notion of a "discourse" in use here to be applied to more coarse-grained parts of speech, such as whole conversations? Answering these questions seems to require

some more specificity with respect to the used notion of a "discourse" and remains important to get a good grasp of how Moorean infelicity ought to be explained.

Possible answer: Though I agree that the notion of a "discourse" is applied pretty loosely, I do not think it prevents us from making theoretical use of it. As I noted in Chapter 2., formulating the Flexible Schema was motivated in part by Goldberg's (2015) observation that in certain domains it is rational to have low expectations of attaining knowledge, which, on the other hand, does not prevent the need to engage in assertoric practice with respect to such domains. I take this point to stand regardless of whether the exact boundaries of such domains are possible to delineate and take the intuitive evaluation in "discourse" terms to capture enough important differences to serve an explanatory role. In my view, a promising approach that would allow for further specification and theoretical refinement of this account, as well as pairing it with the approach to non-assertoric speech acts defended in Chapter 3., would be the one that integrates it with mechanisms structuring the common ground. Roberts (2018) presents an interesting move in this direction, by defining "discourse goals" as sets of questions under discussion that set the ultimate goal of the conversation understood as a common inquiry. Given that the "epistemic prospects" in a given type of inquiry might be limited, an extension of Roberts' proposal would need to take into account the fact that different topic-individuated goals may require different sources or strength of evidence signaled through assertion. See also Yalcin's remarks in his 2007, pp. 1008-1012 on "conversational tone" as a possible indication of a similar view.

Regarding the second, more specific question raised in the *Problem* part, I believe that the natural approach is that each indivisible asserted part of one's utterance ought to be held to the standard specific to its own discourse/topic. In conjunctive cases, this would mean that both conjuncts are evaluated from the point of view of their respective discourses/topics, as both conjuncts were (consecutively) asserted. For Moorean conjunctions, this would mean, that each conjunct is typically held to a different evidential standard; e.g., the assertion "Labour will win the next election, but

I don't know that" should count as warranted if the first conjunct meets FNA-based standard of propriety and the second – AvNA-based or a similar one[148]. To achieve this fine-graininess of conversational standards, one would need to, however, allow for far-reaching modifications of Roberts' framework mentioned above, as it would require a variety of questions under discussion coming in and going out of force during a conversation; I shall not pursue that goal any further here.

- **Iterated epistemic Moorean conjunctions**

Problem: This puzzle, after the title of Sosa's paper which introduced it (2009), was labeled the problem of "dubious assertions": it seems that assertions of the form "*p*, but I don't know that I know *p*" are infelicitous, similarly to unbelievable IOMP constructions discussed in Chapter 4. But if their infelicity were to be explained along the lines of KNA, this seems to require the use of the *KK* principle, which many supporters of the former reject, so either one of these positions needs to be abandoned. Sosa leaves this dilemma open; Cohen and Comesaña (2013) and Greco (2015a) take this to be an argument for *KK*. Since the position defended in this dissertation includes rejection of *KK* and upholding a strong doxastic norm of assertion as a basis for infelicity judgments for Moorean assertions, counting "dubious assertions" among them threatens the analysis presented here.

Possible answer: In response to this challenge, the supporters of KNA: Benton (2013) and Williamson (2013) argued that such assertions are, in fact, not

[148] Another interesting issue here is whether, if the analysis of knowledge self-ascriptions presented in Chapter 5. is correct, avowals or disavowals of knowledge (and other factive states) should be held to the same standard as avowals or disavowals of belief. I am tempted to say that, if they require "Rylean" knowledge of one's belief's epistemic credentials, they should be actually held to a weaker evidential standard than epistemic certainty. Nevertheless, I leave it to the reader to decide whether they find this approach to be plausible.

impermissible, but only *secondarily improper* or *careless*: the speaker may be in a position to assert them, but in doing so, they signal that they do not believe that they have the relevant authority to do so, which makes them sound "clunky" (see also Palczewski 2014, pp. 317-320). Benton (2013, p. 356) also provides examples in which such constructions can be (according to his intuition) felicitously asserted. I concur with his judgment, and, as I also argue in Chapter 2. that some simple EOMP conjunctions can be felicitously asserted, I see the "dubious conjunctions" as no more puzzling than the others[149].

- **Moore's Paradox and Burge-Buridan Paradox**

Problem: As noted in Chapter 4., the anti-expertise sentence, which I included in the Moorean class, bears a certain similarity to the paradoxical, self-referential sentence of the form "I don't believe that this sentence is true", which was first explicitly discussed by Burge (1978). As he locates its source in Buridan's *Sophismata*, it came to be known as a Burge-Buridan[150] paradoxical sentence (Lenzen 1981). If we consider the puzzling sentence on its own, it seems to meet our tests for Moore-paradoxicality for rational belief just as anti-expertise sentences do (i.e. it is not irrational to believe it's third-person or past tense counterpart or to suppose it). Following Montague (1963) and Lenzen, I said that the primary lesson to be learned from it is that "belief" should not be formalized as a sentential predicate, but as an operator, which aligned with my use of modal doxastic logic as the formal model for the logic of belief; in such settings, Burge-Buridan sentences are simply not constructible. However, this move can be seen

[149] Another option consistent with my account is that in cases in which "dubious" assertions really are indefensible, a particularly strong discourse-specific norm is in place, e.g. AvNA.

[150] It is also sometimes referred to as "Godel sentences", after Smullyan 1986, who informally likens it to the sentence involved in the proof of Godel's Second Incompleteness Theorem.

as dubitable, if we consider that Burge-Buridan sentences *are* constructible in natural language and, if we grant that they have truth conditions, ostensively true only if not believed, and *this* fact presumably should also merit some discussion.

Possible answer: To analyze Burge-Buridan's paradoxical sentence in a formal setting, one needs more advanced formal devices than is covered in this dissertation. If one introduces "belief" as a sentential predicate to a sufficiently strong basic theory (e.g. Peano Arithmetic), and the Burge-Buridan sentence becomes constructible and provably equivalent to its unbelievability *via* the fixed-point lemma, one may show that such theory is classically inconsistent if OMP-sentences are taken to be unbelievable in the theory (Schuster, Horsten 2022). An intuitive idea would be to model such theories non-classically, in a way analogous to certain formal approaches to the Liar Paradox; for a theory of rational belief and knowledge analyzed in such a way that evades the unwelcome result by mirroring Kripke's non-classical semantics for truth, see Schuster 2023. The downside of this treatment is that, while it maintains that the Burge-Buridan sentence is not rationally believed in Schuster's model, it is also not *not believed* (due to the fact that the proposed model does not differentiate between belief suspension and disbelief), and is neither true nor false (Schuster 2023, pp. 15-16). While a broadly philosophical justification of this fact could be perhaps given in terms of this sentence being, in Kripke's words, *ungrounded* and therefore *inapt* for being an object of belief, it clearly remains problematic (as Caie [2013] also notices of similar systems, arguing against non-classical approaches to Burge-Buridan-style paradoxes). From the perspective of this dissertation, it would be desirable to have such a non-classical theory for belief predicate that allows to hold that the Burge-Buridan sentence *cannot be believed* and yet *is provably true*, which may very well be impossible. Given, however, that a formal study of the problem is still not fully developed, I will suspend my judgment on the matter.

6.3. Conclusion

Overall, one can bring down the analysis provided in this dissertation to two simple constatations: assertion is minimally constrained by what one may rationally believe, and what one may rationally believe is determined in part by what one can come to know about one's own beliefs. In a way, neither of these claims is especially groundbreaking, and both were defended in many different flavors by a plurality of philosophers with widely different inclinations. The main aim of this dissertation was to defend these claims against objections (supplementing, explicating, rephrasing or qualifying them if necessary) and demonstrate that upholding them is plausibly the *only* way we can hope to arrive at a uniform solution to Moore's Paradox in both speech and thought.

If this dissertation represented some program in the philosophy of language and epistemology, it would be, no doubt, on the conservative side. The main three inspirations I drew from can be easily identified as now-classic writings of Tim Williamson, Bob Stalnaker, and Jaakko Hintikka. Though I disagreed with Williamson with respect to the scope of both his knowledge account of assertion and critique of luminosity, I take both Chapters 2. and 5. to be broadly Williamsonian in spirit; similarly, I think that Stalnaker's work in pragmatics and formal epistemology inspired both parts of this dissertation, besides the obvious debt I owe to his analysis in Chapter 3. Hintikka, as the founding father of formal epistemology and epistemic and doxastic logic, obviously inspired the methods of investigation and the argument made in Chapter 4., but upon rereading *Knowledge and Belief* I found more and more inspiring passages that influenced my thinking beyond that – as may be, I think, evidently seen in Chapter 2.

After acknowledging these influences, let me finish this dissertation with a story. As is told to many students of mathematics starting to learn set theory, Georg Cantor once commented in a letter to Richard Dedekind on one of his theorems: *I see it, but I don't believe it*. These words are commonly interpreted as Moore-paradoxically

sounding evidence that what one proves in set theory may be surprising beyond belief, yet one should always remain epistemically humble and follow the proof. The truth about the circumstances of Cantor's utterance is, however, more interesting. When put in a context[151], Cantor's words were not meant to express surprise, but a trust in his epistemic peer; what Cantor meant was that, though he was certain of his result, he cautiously awaited Dedekind's opinion on its correctness to fully believe it. As my work here concludes, I similarly put my trust in the judgment of the reader.

[151] The full quote goes: „I can have no peace of mind until I obtain from you, honoured friend, a decision about [my result's] correctness. So long as you have not agreed with me, I can only say: *je le vois, mais je ne le crois pas* [I see it, but I don't believe it]. And so I ask you to send me a postcard and let me know when you expect to have examined the matter, and whether I can count on an answer to my quite demanding request." (after: Gouvêa 2011, p. 207).

References

1. Adler, J. E. (2002). *Belief's own ethics*. Cambridge, MA: MIT Press.
2. Ajdukiewicz, K. (1967). Proposition as the Connotation of Sentence. *Studia Logica*, *20*, 87-98.
3. Alston, W. P. (2000). *Illocutionary Acts and Sentence Meaning*, Ithaca, NY: Cornell University Press.
4. Antony, L. (2004). A Naturalized Approach to the "A Priori". *Philosophical Issues*, *14*, 1-17.
5. Apter, A. (2017). Ethnographic X-files and Holbraad's double-bind: Reflections on an ontological turn of events. *HAU: Journal of Ethnographic Theory*, *7*(1), 287-302.
6. Ariel, M. (2010). *Defining pragmatics*. Cambridge: Cambridge University Press.
7. Armstrong, D. (1968). *A Materialist Theory of the Mind*. London: Routledge.
8. Atlas, J. D. (2005). *Logic, meaning, and conversation: Semantical underdeterminacy, implicature, and their interface*. Oxford: Oxford University Press.
9. Austin, J. L. (1940/1979). The meaning of a word. In J. O. Urmson, G. J. Warnock (Eds.) *Philosophical Papers*. Oxford: Clarendon Press. (Published posthumously).
10. Austin, J. L. (1962). *How to do things with words*. Oxford: Clarendon Press.
11. Bach, K. (2006). The top 10 misconceptions about implicature. In R. Horn, B. Birner, & G. Ward, (Eds.), *Drawing the boundaries of meaning. Neo-Gricean studies in pragmatics and semantics in honor of Laurence*. Amsterdam: John Benjamins.
12. Bach, K. (2008). Applying Pragmatics to Epistemology, *Philosophical Issues*, 18: 68–88.
13. Bach, K., Harnish, R. M. (1979). *Communication and speech acts*. Cambridge, MA: Harvard University Press.

14. Bach, K., Harnish, R. M. (1992). How performatives really work: A reply to Searle. *Linguistics and Philosophy*, *15*(1), 93-110.
15. Baldwin, T. (1990). *G.E. Moore*. New York: Routledge.
16. Bar-Hillel, Y. (1946). Analysis of "correct" language. *Mind*, *55*(220), 328-340.
17. Bar-Hillel, Y. (1954). Indexical expressions. *Mind*, *63*(251), 359-379.
18. Bar-Hillel, Y. (1971). Out of the pragmatic wastebasket. *Linguistic Inquiry*, *2*(3), 401-407.
19. Bar-On, D. (2004). *Speaking my mind: Expression and self-knowledge*. Oxford: Clarendon Press.
20. Bartha, K. (2021). Monolingual and bilingual children's understanding of Moore-paradox sentences. *Cognition, Brain, Behavior. An Interdisciplinary Journal*, *25*(2), 129-155.
21. Benton, M. A. (2012). Assertion, knowledge and predictions. *Analysis*, *72*(1), 102-105.
22. Benton, M. A. (2013). Dubious objections from iterated conjunctions. *Philosophical Studies*, *162*, 355-358.
23. Benton, M. A. (2016). Gricean quality. *Noûs*, *50*(4), 689-703.
24. Bird, A., & Pettigrew, R. (2021). Internalism, externalism, and the KK principle. *Erkenntnis*, *86*, 1713-1732.
25. Black, M. (1952). Saying and disbelieving. *Analysis*, *13*(2), 25-33.
26. Borgoni, C. (2015). Dissonance and Moorean propositions. *dialectica*, *69*(1), 107-127.
27. Bortolotti, L. (2004). Can we interpret irrational behavior?. *Behavior and Philosophy*, 359-375.
28. Boyle, M. (2011). Transparent Self-Knowledge. *Aristotelian Society Supplementary Volume* Vol. 85, No. 1, 223-241.
29. Brandom, R. (1983). Asserting. *Noûs*, 637-650.

30. Bromberger, S. (1966). *Why-questions*. In R. G. Colodny (Ed.) (1966). *Mind and Cosmos: Essays in Contemporary Science and Philosophy*. Pittsburgh: Pittsburgh University Press.
31. Burge, T. (1978). Buridan and epistemic paradox. *Philosophical Studies*, *34*(1), 21-35.
32. Burge, T. (1996). Our entitlement to self-knowledge. *Proceedings of the Aristotelian Society*, Vol. 96, 117-158.
33. Byrne, A. (2005). Introspection. *Philosophical topics*, *33*(1), 79-104.
34. Byrne, A. (2011). Knowing that I am thinking. In A. Hatzimoysis (Ed.) (2011). *Self-knowledge*. Oxford: Oxford University Press.
35. Byrne, A. (2012). Knowing what I see. In Smithies, D., Stoljar, D. (Eds.). (2012). *Introspection and consciousness*. Oxford: Oxford University Press.
36. Byrne, A. (2018). *Transparency and self-knowledge*. Oxford: Oxford University Press.
37. Byrne, A. (2021). Perception and probability. *Philosophy and Phenomenological Research*, 104, 343–363.
38. Caie, M. (2013). Belief and indeterminacy. *The Philosophical Review, 122*(4), 527–575.
39. Cappelen, H. (2011). Against assertion. In J. Brown, H. Cappelen (Eds.) (2011). *Assertion: New philosophical essays*. Oxford: Oxford University Press.
40. Carnap, R. (1988). *Meaning and necessity: A study in semantics and modal logic*. Chicago: University of Chicago Press.
41. Carston, R. (2002). *Thoughts and utterances: The pragmatics of explicit communication*. Oxford: Blackwell.
42. Castañeda, H. N. (1970). On knowing (or believing) that one knows (or believes). *Synthese*, 187-203.
43. Chan, T. (2010). Moore's paradox is not just another pragmatic paradox. *Synthese*, *173*(3), 211-229.

44. Chellas, B. F. (1980). *Modal Logic: An Introduction*. Cambridge: Cambridge University Press.
45. Cholbi, M. (2009). Moore's paradox and moral motivation. *Ethical theory and moral practice*, *12*, 495-510.
46. Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.
47. Christensen, D. (2010). Higher-order evidence. *Philosophy and Phenomenological Research, 81*(1), 185–215.
48. Ciecierski, T. (2023). A note on the demonstrative uses of indexicals. *Logique et Analyse*, *258*, 151-166.
49. Clapp, L. (2020). Stalnaker on the essential effect of assertion. In: Goldberg, S.C. (ed.) (2020). *The Oxford Handbook of Assertion*, Oxford: Oxford University Press.
50. Clark, M. (2012). *Paradoxes from A to Z*. London: Routledge.
51. Cohen, L. J. (1964). Do illocutionary forces exist?. *The Philosophical Quarterly 14*, 420–444.
52. Cohen, S., & Comesaña, J. (2013). Williamson on Gettier cases and epistemic logic. *Inquiry*, *56*(1), 15-29.
53. Coliva, A. (2015). How to commit Moore's paradox. *The Journal of Philosophy*, *112*(4), 169-192.
54. Collins, J. (2021). A norm of aesthetic assertion and its semantic (in) significance. *Inquiry*, *64*(10), 973-1003.
55. Condoravdi, C., Lauer, S. (2012). Imperatives: Meaning and illocutionary force. *Empirical Issues in Syntax and Semantics, 9*, 37–58.
56. Conee, E. (1982). Utilitarianism and rationality. *Analysis, 42*(1), 55–59.
57. Corazza, E., Fish, B., Gorvett, J. (2002). Who is I? *Philosophical Studies, 107*, 1–21.
58. Crimmins, M. (1992). I falsely believe that p. *Analysis 52*, 191.
59. Crimmins, M., Perry, J. (1989). The prince and the phone booth: Reporting puzzling beliefs. *The Journal of Philosophy*, *86*(12), 685-711.
60. Das, N., Salow, B. (2018). Transparency and the KK Principle. *Noûs*, *52*(1), 3-23.

61. DeRose, K. (1991). Epistemic possibilities. *The Philosophical Review*, *100*(4), 581-605.
62. DeRose, K. (2002). Assertion, knowledge, and context. *The Philosophical Review*, *111*(2), 167-203.
63. Dinges, A. (2023). Assertion and certainty. *The Philosophical Quarterly*, pqad022.
64. Dokic, J., & Égré, P. (2009). Margin for error and the transparency of knowledge. *Synthese*, *166*, 1-20.
65. Dorst, K., Mandelkern, M. (2022). Good guesses. *Philosophy and Phenomenological Research*, *105*(3), 581-618.
66. Douven, I. (2006). Assertion, knowledge, and rational credibility. *The Philosophical Review*, *115*(4), 449-485.
67. Douven, I. (2009). Assertion, Moore, and Bayes. *Philosophical Studies*, *144*, 361-375.
68. Egan, A., Elga, A. (2005). I can't believe I'm stupid. *Philosophical Perspectives*, *19*, 77-93.
69. Evans, G. (1982). *Varieties of Reference.* Oxford: Clarendon Press.
70. Fagin, R., Halpern, J. Y., Moses, Y., Vardi, M. (2004). *Reasoning about knowledge*. Cambridge, MA: MIT Press.
71. Faller, M. (2002) *Semantics and Pragmatics of Evidentials in Cuzco Quechua*. PhD thesis, Stanford University.
72. Fernández, J. (2013). *Transparent minds: A study of self-knowledge*. Oxford: Oxford University Press.
73. Fileva, I., Brakel, L. A. (2019). Just another article on Moore's paradox, but we don't believe that. *Synthese*, *196*(12), 5153-5167.
74. Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Oxford: Oxford University Press.
75. Frances, B. (2016). Rationally held 'P, but I fully believe ~P and I am not equivocating'. *Philosophical Studies*, *173*(2), 309-313.

76. Frege, G. (1918/1956). The thought: A logical inquiry. (Trans. A. M. Quinton, M. Quinton), *Mind*, *65*(259), 289-311 (Original work published 1918).
77. Gallois, A. (1996). *The world without, the mind within: An essay on first-person authority*. Cambridge: Cambridge University Press.
78. García-Carpintero, M. (2013). Explicit performatives revisited. *Journal of Pragmatics*, *49*(1), 1-17.
79. Gaszczyk, G. (2022). Norms of speech acts. *Studia Semiotyczne*, *36*(2), 11-45.
80. Gaszczyk, G. (2023). Helping others to understand: A normative account of the speech act of explanation. *Topoi*, *42*(2), 385-396.
81. Geach, P. T. (1965). Assertion. *The Philosophical Review*, *74*(4), 449-465.
82. Gerken, M. (2012). Discursive justification and skepticism. *Synthese*, *189*(2), 373-394.
83. Gertler, B. (2012). Renewed acquaintance. In D. Smithies, D. Stoljar, (Eds.). (2012). *Introspection and consciousness,* 89-123. Oxford: Oxford University Press.
84. Gertler, B. (2021a). Self-Knowledge. In E. N. Zalta, *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition). URL = <https://plato.stanford.edu/archives/win2021/entries/self-knowledge/>.
85. Gertler, B. (2021b). Smithies on Self-Knowledge of Beliefs. *Analysis*, *81*(4), 782-792.
86. Ginet, C. (1979). Performativity. *Linguistics and Philosophy*, *3*(2), 245-265.
87. Goldberg, S. (2015). *Assertion: On the philosophical significance of assertoric speech*. Oxford: Oxford University Press.
88. Goldman, A. I. (1986). *Epistemology and Cognition*, Cambridge, MA: Harvard University Press.
89. Goldstein, L. (1988). Wittgenstein's Late Views on Belief, Paradox and Contradiction. *Philosophical Investigations*, *11*(1), 49-73.
90. Goldstein, L. (1993). Inescapable Surprises and Acquirable Intentions. *Analysis, 53*(2): 93–9.
91. Goldstein, S. (2024). *Iterated Knowledge.* Oxford: Oxford University Press.

92. Gordon, E. C. (2023). Understanding of the norm of political discourse. *Synthese*, *201*(6), 1-13.

93. Gouvêa, F. Q. (2011). Was Cantor surprised?. *The American Mathematical Monthly*, *118*(3), 198-209.

94. Grant, C. K. (1958). Pragmatic Implication. *Philosophy*, *33*(127), 303-324.

95. Greco, D. (2014). Could KK be ok?. *The Journal of Philosophy*, *111*(4), 169-197.

96. Greco, D. (2015a). Iteration principles in epistemology I: Arguments for. *Philosophy Compass*, *10*(11), 754-764.

97. Greco, D. (2015b). How I learned to stop worrying and love probability 1. *Philosophical Perspectives*, *29*, 179-201.

98. Green, M. S. (2007). Moorean absurdity and showing what's within. In M. S. Green, J. N. Williams (Eds.), *Moore's Paradox: New essays on belief, rationality, and the first person*. Oxford: Clarendon Press.

99. Green, M. S. (2020). Assertion and Convention. In: Goldberg, S.C. (ed.) (2020). *The Oxford Handbook of Assertion*, Oxford: Oxford University Press.

100. Grice, H. P. (1971). Intention and Uncertainty. *The Proceedings of the British Academy, Vol. LVII*, 3-19.

101. Grice, H. P. (1975/1989). Logic and Conversation. In H.P. Grice (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

102. Grice, H. P. (1978/1989). Further notes on logic and conversation. In H.P. Grice (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

103. Grice, H. P. (1981). Presupposition and conversational implicature. In P. Cole (Ed.), *Radical pragmatics*. New York: Academic Press.

104. Hacquard, V. (2006). *Aspects of modality*. PhD thesis, Massachusetts Institute of Technology.

105. Hájek, A. (2007). My philosophical position says ‘p’ and I don’t believe ‘p’. In: M. S. Green, J. N. Williams (eds.), *Moore’s Paradox: New Essays on Belief, Rationality and the First Person*, Oxford: Clarendon Press.

106. Harris, D. W., Fogal, D., Moss, M. (2018). Speech acts: The contemporary theoretical landscape. In: D. Fogal, D. W. Harris, M. Moss (Eds.) (2018). *New work on speech acts*. Oxford: Oxford University Press.

107. Harsanyi, J. (1967). Games of Incomplete Information Played by Bayesian Players. Parts I, II, III. *Management Science 14*, 159-182, 320-334, 486-502.

108. Haslanger, S. (1992). Ontology and pragmatic paradox. *Proceedings of the Aristotelian Society*, 92, 293-313.

109. Hawthorne, J., Magidor, O. (2009). Assertion, context, and epistemic accessibility. *Mind*, *118*(470), 377-397.

110. Hawthorne, J., Rothschild, D., Spectre, L. (2016). Belief is weak. *Philosophical Studies*, *173*, 1393-1404.

111. Heal, J. (1978). Common knowledge. *The Philosophical Quarterly, 28*(111), 116-131.

112. Heal, J. (1994). Moore's paradox: A Wittgensteinian approach. *Mind*, *103*(409), 5-24.

113. Hindriks, F. (2007). The Status of the Knowledge Account of Assertion. *Linguistics and Philosophy*, 30(3): 393–406.

114. Hintikka, J. (1962). *Knowledge and belief: An introduction to the logic of the two notions.* Ithaca: Cornell University Press.

115. Holliday, W. (2018). Epistemic Logic and Epistemology. In S. O. Hansson, V. F. Hendricks (eds.) (2018). *Introduction to Formal Philosophy*. Cham: Springer.

116. Hong, F. (2021). Uttering Moorean Sentences and the pragmatics of belief reports. *Philosophical Studies*, *178*, 1879-1895.

117. Huemer, M. (2007). Moore's Paradox and the Norm of Belief. In S. Nuccetelli, G. Seay (Eds.) (2007). *Themes from G. E. Moore: New essays in epistemology and ethics*, Oxford: Oxford University Press.

118. Huemer, M. (2011). The puzzle of metacoherence. *Philosophy and phenomenological research*, *82*(1), 1-21.

119. Hume, D. (1739/1896). *Treatise on Human Nature.* Oxford: Clarendon Press.

120. Ichikawa, J. J., Steup, M. (2018). The Analysis of Knowledge, In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), URL = <https://plato.stanford.edu/archives/sum2018/entries/knowledge-analysis/>.

121. Jago, M. (2014). *The impossible: An essay on hyperintensionality*. Oxford: Oxford University Press.

122. James, W. (1890). *The Principles of Psychology, Volume 1.* New York: Henry Holt and Company.

123. Jary, M. (2007). Are explicit performatives assertions?. *Linguistics and Philosophy*, *30*, 207-234.

124. Kaplan, D. (1989), *Demonstratives*. In J. Almog, J. Perry, H. Wettstein (Eds.) (1989). *Themes from Kaplan*. Oxford: Oxford University Press.

125. Kelp, C. (2020). Moral Assertion. *Ethical Theory and Moral Practice*, 23(3–4), 639–649.

126. King, J. C., Soames, S., & Speaks, J. (2014). *New thinking about propositions*. Oxford: Oxford University Press.

127. Kriegel, U. (2004). Moore's paradox and the structure of conscious belief. *Erkenntnis*, *61*(1), 99-121.

128. Kripke, S. (1979). A Puzzle about belief. In A. Margalit (Ed.) (1979). *Meaning and Use*, Dordrecht: Springer.

129. Kripke, S. (2011). *Philosophical Troubles: Collected Papers, Volume 1*. Oxford: Oxford University Press.

130. Kroon, F. W. (1990). On a Moorean solution to instability puzzles. *Australasian Journal of Philosophy*, *68*(4), 455-461.

131. Kvanvig, J. (2009). Assertion, Knowledge, and Lotteries. In D. Pritchard, P. Greenough (eds.) (2009). *Williamson on Knowledge*. Oxford: Oxford University Press.

132. Lackey, J. (2007). Norms of assertion. *Noûs*, *41*(4), 594-626.

133. Langford, C. H. (1942). The Notion of Analysis in Moore's Philosophy. In P. A. Schlipp (Ed.) (1942). *The Philosophy of George Edward Moore*, Menasha: George Banta Publishing Company.

134. Lasonen-Aarnio, M. (2008). Single premise deduction and risk. *Philosophical Studies*, 141, 157-173.

135. Lauer, S. (2014). Mandatory implicatures in Gricean pragmatics. In J. Degen, M. Franke, & N. Goodman (Eds.) (2014). *Proceedings of the formal & experimental pragmatics workshop*, Tübingen.

136. Lechniak, M. (2011). *Przekonania i zmiana przekonań: analiza logiczna i filozoficzna.* Lublin: Wydawnictwo KUL.

137. Lechniak, M. (2018). Once More about Moore's Paradox in Epistemic Logic and Belief Change Theory. *Roczniki Filozoficzne*, *66*(3), 77-99.

138. Lederman, H. (2018). Uncommon knowledge. *Mind*, *127*(508), 1069-1105.

139. Lemmon, E. (1962). On sentences verifiable by their use. *Analysis*, *22*(4), 86-89.

140. Lenzen, W. (1979). Epistemologische betrachtungen zu [s4, s5]. *Erkenntnis*, *14*, 33-56.

141. Lenzen, W. (1981). Doxastic logic and the Burge-Buridan-paradox. *Philosophical Studies*, 43-49.

142. Levi, I. (1991). *The fixation of belief and its undoing: Changing beliefs through inquiry*. Cambridge: Cambridge University Press.

143. Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

144. Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, *8*, 339-359.

145. Linsky, B. (1986). Factives, blindspots and some paradoxes. *Analysis*, *46*(1), 10-15.

146. Linville, K., Ring, M. (1991). Moore's paradox revisited. In J. Hintikka (Ed.), *Wittgenstein in Florida: Proceedings of the Colloquium on the Philosophy of Ludwig Wittgenstein, Florida State University, 7–8 August 1989*. Dordrecht: Springer.

147. Littlejohn, C. (2010). Moore's paradox and epistemic norms. *Australasian Journal of Philosophy*, *88*(1), 79-100.

148. Littlejohn, C. (2020). Moore's paradox and assertion. In S. C. Goldberg (Ed.), *Oxford handbook of assertion*. Oxford: Oxford University Press.

149. Łoś, J. (1948). Logiki wielowartościowe a formalizacja funkcji intensjonalnych. *Kwartalnik Filozoficzny*, *17*(1), 59-78.

150. Lycan, W. (1996). *Consciousness and Experience*. Cambridge, MA: MIT Press.

151. Lycan, W. (2010). What, exactly, is a paradox?. *Analysis*, *70*(4), 615-622.

152. MacDonald, M. (1937). Induction and Hypothesis. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *16*, 20-102.

153. MacFarlane, J. (unpublished manuscript). Epistemic Modalities and Relative Truth. Available online at: https://johnmacfarlane.net/epistmod-2003.pdf.

154. MacIver, A. M. (1938). Some Questions about" Know" and "Think". *Analysis*, *5*(3/4), 43-50.

155. Maitra, I. (2011). Assertion, Norms and Games. In J. Brown and H. Cappelen (Eds.), *Assertion: New Philosophical Essays*. Oxford: Oxford University Press.

156. Maitra, I., Weatherson, B. (2010). Assertion, knowledge, and action. *Philosophical Studies*, *149*, 99-118.

157. Malcolm, N. (1950/1975). The Verification Argument. In N. Malcolm, *Knowledge and Certainty: Essays and Lectures*. London: Prentice Hall, Inc.

158. Malcolm, N. (1995). Disentangling Moore's Paradox. In R. Egidi (Ed.) *Wittgenstein: Mind and Language*. Springer: Dordrecht.

159. Mandelkern, M. (2019). Bounded modality. *The Philosophical Review*, *128*(1), 1-61.

160. Mandelkern, M. (2021). Practical Moore sentences. *Noûs*, *55*(1), 39-61.

161. Mandelkern, M., Dorst, K. (2022). Assertion is weak. *Philosophers' Imprint* 22: 19.

162. Marciszewski, W. (1972). *Podstawy logicznej teorii przekonań.* Warszawa: PWN.

163. Martinich, A. P. (1980). Conversational maxims and some philosophical problems. *The Philosophical Quarterly (1950-)*, *30*(120), 215-228.

164. McCready, E. (2015). *Reliability in pragmatics*. Oxford: Oxford University Press.

165. McGlynn, A. (2013). Believing things unknown. *Noûs*, *47*(2), 385-407.

166. McGlynn, A. (2014). *Knowledge first?*. London: Palgrave Macmillan.

167. McGuinness, B. (2008). *Wittgenstein in Cambridge: letters and documents, 1911-1951.* Oxford: Blackwell.

168. McHugh, C. (2010). Self-knowledge and the KK principle. *Synthese*, *173*(3), 231-257.

169. McKinnon, R. (2013). The Supportive Reasons Norm of Assertion. *American Philosophical Quarterly*, 50: 121–35.

170. McKinnon, R. (2015). *The Norms of Assertion: Truth, Lies, and Warrant*. New York*:* Palgrave Macmillan.

171. Monk, R. (1991). *Ludwig Wittgenstein: The Duty of Genius*. New York: Vintage Books.

172. Montague, R. (1963). Syntactical Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability, *Acta Philosophica Fennica*, *Vol. 16*, 153-167.

173. Montminy, M. (2013). Why assertion and practical reasoning must be governed by the same epistemic norm. *Pacific Philosophical Quarterly*, *94*(1), 57-68.

174. Moore G. E. (1903). The refutation of idealism. *Mind*, *12*(48), 433-453.

175. Moore, G. E. (1907). *Ethics*. New York: Henry Holt and Company.

176. Moore, G. E. (1942). Reply to My Critics. In P. A. Schlipp (Ed.) (1942). *The Philosophy of George Edward Moore*, Menasha: George Banta Publishing Company.

177. Moore, G. E. (1944). Bertrand Russell's Theory of Descriptions. In P. A. Schlipp (Ed.) (1944). *The Philosophy of Bertrand Russell*, Menasha: George Banta Publishing Company.

178. Moore, G. E. (1962). *Commonplace Book 1919-1953*. London: George Allen & Unwin Ltd.

179. Moore, G. E. (1993). Moore's Paradox. In T. Baldwin (Ed.) (1993). *G.E. Moore: Selected Writings*, London: Routledge.

180. Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge*. Princeton: Princeton University Press.

181. Nichols, S., Stich, S. (2003). *Mindreading: An Integrated Account of Pretence, Awareness, and Understanding Other Minds*. Oxford: Oxford University Press.

182. Ninan, D. (2005). Two puzzles about deontic necessity. In: J. Gajewski, V. Hacquard, B. Nickel, and S. Yalcin (Eds.) (2005). *New Work on Modality. MIT Working Papers in Linguistics.*

183. Ninan, D. (2014). Taste predicates and the acquaintance inference. In T. Snider, S. D'Antonio, M. Weigand (Eds.). (2014). *Semantics and Linguistic Theory (SALT) 24*. Ithaca, NY: CLC.

184. O'Connor, D. J. (1948). Pragmatic paradoxes. *Mind, 57*, 358–359.

185. Okasha, S. (2013). On a flawed argument against the KK principle. *Analysis*, *73*(1), 80-86.

186. Pagin, P. (2016). Problems with norms of assertion. *Philosophy and Phenomenological Research*, *93*(1), 178-207.

187. Pagin, P., Marsili, N. (2021). Assertion. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), URL = <https://plato.stanford.edu/archives/win2021/entries/assertion/>.

188. Palczewski, R. (2014). *Wiedza w kontekstach: w obronie kontekstualizmu epistemicznego. Tom I: Między pragmatyką a semantyką.* Toruń: Wydawnictwo UMK.

189. Peluce, V. A. (2017). From epistemic paradox to doxastic arithmetic. In S. Artemov, S. Nerode (Eds.), *International Symposium on Logical Foundations of Computer Science*. Cham: Springer.

190. Perry, J. (1979). The problem of the essential indexical. *Noûs*, 3-21.

191. Pettit, P. (1987). Humeans, anti-Humeans, and motivation. *Mind*, *96*(384), 530-533.

192. Portner, P. (2004). The semantics of imperatives within a theory of clause types. In R. Young (ed.), *Semantics and Linguistic Theory (SALT) XIV,* Ithaca, NY: Cornell University Press.

193. Portner, P. (2007). Imperatives and modals. *Natural language semantics*, *15*, 351-383.

194. Predelli, S. (1998). 'I am not here now'. *Analysis*, *58*(2), 107-115.

195. Pruss, A. R. (2012). Sincerely asserting what you do not believe. *Australasian Journal of Philosophy*, *90*(3), 541-546.

196. Puczyłowski, T. A. (2020). Odwoływalność i mówienie nie wprost. *Filozofia Nauki*, *28*(3 (111)), 73-98.

197. Quine, W. V. (1962). Paradox. *Scientific American*, *206*(4), 84-99.

198. Ramachandran, M. (2012). The kk-principle, margins for error, and safety. *Erkenntnis*, *76*, 121-136.

199. Rawls, J. (1955). Two concepts of rules. *The Philosophical Review*, *64*(1), 3-32.

200. Rescher, N. (1968). *Topics in Philosophical Logic.* Dordrecht: Reidel.

201. Richter, R. (1990). Ideal rationality and hand waving. *Australasian Journal of Philosophy*, 68(2), 147-156.

202. Rieger, A. (2015). Moore's paradox, introspection and doxastic logic. *Thought: A Journal of Philosophy*, *4*(4), 215-227.

203. Ring, M., Linville, K. (1973). Moore's Paradox: Assertion and Implication. *Behaviorism*, *1*(2), 87-102.

204. Roberts, C. (2018). Speech acts in discourse context. In: D. Fogal, D. W. Harris, M. Moss (Eds.) (2018). *New work on speech acts*. Oxford: Oxford University Press.

205. Russell, B. (1912). *Problems of Philosophy*. New York: Henry Holt and Company.

206. Russell, B. (1921). *The Analysis of Mind.* London: George Allen & Unwin Ltd.

207. Russell, B. (1926). Theory of Knowledge. In: *Encyclopaedia Britannica*, Thirteenth Edition, Volume II, London: The Encyclopaedia Britannica Company.

208. Ryle G. (1949). *The Concept of Mind*. Chicago: University of Chicago Press.

209. Sadock, J. (1978). On testing for conversational implicature. In P. Cole (Ed.), *Syntax and Semantics: Pragmatics.* New York: Academic Press.

210. Sainsbury, R. M. (2009). *Paradoxes*. Cambridge: Cambridge University Press.

211. Salmon, N. (1986). *Frege's Puzzle*. Cambridge, MA: MIT Press.

212. San, W. K. (2020). Fitch's Paradox and Level-Bridging Principles. *The Journal of Philosophy*, *117*(1), 5-29.

213. Schuster, D. (2023). The fixed points of belief and knowledge. *Logic Journal of the IGPL*, jzad016.

214. Schuster, D., Horsten, L. (2022). On the pure logic of justified belief. *Synthese*, *200*(5), 425.

215. Schütze, T. C. (2016). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology, 2nd edition*. Berlin: Language Science Press.

216. Searle, J. (1970). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

217. Searle, J. (1976). A classification of illocutionary acts. *Language in society*, *5*(1), 1-23.

218. Searle, J. (1989). How performatives work. *Linguistics and philosophy*, *12*, 535-558.

219. Searle, J. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

220. Searle, J., Vanderveken, D. (1985). *Foundations of illocutionary logic*. Cambridge: Cambridge University Press.

221. Semeijn, M. (2021). *Fiction and common ground: a workspace account*. PhD thesis, University of Groningen.

222. Serota, K. B., Levine, T. R. (2015). A few prolific liars: Variation in the prevalence of lying. *Journal of Language and Social Psychology*, *34*(2), 138-157.

223. Shah, N., Velleman, J. D. (2005). Doxastic deliberation. *The Philosophical Review*, *114*(4), 497-534.

224. Shoemaker, S. (1988). On knowing one's own mind. *Philosophical perspectives*, *2*, 183-209.

225. Shoemaker, S. (1995). Moore's paradox and self-knowledge. *Philosophical Studies*, *77*(2/3), 211-228.

226. Silins, N. (2012). Judgment as a Guide to Belief. In D. Smithies, D. Stoljar (Eds.) (2012). *Introspection and consciousness*, 295-328. Oxford: Oxford University Press.

227. Simion, M., Kelp, C. (2020). Assertion: The Constitutive Norms View. In: Goldberg, S.C. (ed.) (2020). *The Oxford Handbook of Assertion*, Oxford: Oxford University Press.

228. Smithies, D. (2015). Ideal rationality and logical omniscience. *Synthese*, *192*, 2769-2793.

229. Smithies, D. (2016). Belief and Self-Knowledge. *Philosophical Issues*, *26*, 393-421.

230. Smithies, D. (2019). *The epistemic role of consciousness*. Oxford: Oxford University Press.

231. Smullyan, R. (1986). Logicians who reason about themselves. In J. Y. Halpern (Ed.) (1986). *Theoretical aspects of reasoning about knowledge*. Los Altos: Morgan Kaufmann Publishers.

232. Sorensen, R. (1987). Anti-expertise, instability, and rational choice. *Australasian Journal of Philosophy*, *65*(3), 301-315.

233. Sorensen, R. (1988). *Blindspots*. Oxford: Clarendon Press.

234. Sorensen, R. (2000). Moore's problem with iterated belief. *The Philosophical Quarterly*, *50*(198), 28-43.

235. Sorensen, R. (2007). The All-Seeing Eye: A Blind Spot in the History of Ideas. In M. S. Green, J. N. Williams (Eds.) (2007). *Moore's Paradox: New essays on belief, rationality, and the first person*. Oxford: Clarendon Press.

236. Sosa, D. (2009). Dubious assertions. *Philosophical Studies*, *146*(2), 269-272.

237. Sosa, E. (1999). How to defeat opposition to Moore. *Philosophical perspectives*, *13*, 141-153.

238. Sosa, E. (2011). *Knowing Full Well.* Princeton, NJ: Princeton University Press.

239. Stalnaker, R. (1973). Presuppositions. *Journal of Philosophical Logic*, 447-457.

240. Stalnaker, R. (1974). Pragmatic presuppositions. In M. Munitz, P. Unger (eds.) (1974). *Semantics and Philosophy*. New York: New York University Press.

241. Stalnaker, R. (1978). Assertion. In P. Cole (Ed.) (1978). *Pragmatics*. Leiden: Brill.

242. Stalnaker, R. (1984). *Inquiry*. Cambridge: Cambridge University Press.

243. Stalnaker, R. (2002). Common ground. *Linguistics and philosophy*, *25*(5/6), 701-721.

244. Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies*, *128*(1), 169-199.

245. Stanley, J. (2008). Knowledge and certainty. *Philosophical Issues*, *18*, 35-57.

246. Strawson, P. F. (1962). Freedom and Resentment. *Proceedings of the British Academy*, *48,* 1-25.

247. Stroud, B. (2000). *Understanding human knowledge: Philosophical essays*. New York: Oxford University Press.

248. Tarnowski, M. (2022). Wiedza bezpośrednia a przejrzystość introspekcji. *Przegląd Filozoficzny – Nowa Seria, 4*(124), 279-297.

249. Tarnowski, M. (2023). My religion preaches 'p', but I don't believe that p: Moore's Paradox in religious assertions. *Religious Studies*, *First View,* 1-16.

250. Tarnowski, M. (2024). Knowing What One Likes: Epistemicist Solution to Faultless Disagreement. *Acta Analytica*, *OnlineFirst,* 1-20.

251. Tarnowski, M., Głowacki, M. (2022). Words on Kripke's Puzzle. *Synthese*, *200*(4), 292.

252. Tennessen, H. (1959). Logical oddities and locutional scarcities: Another attack upon methods of revelation. *Synthese*, *11*, 369-388.

253. Tennessen, H. (1961). Whereof one has been silent, thereof one may have to speak. *The Journal of Philosophy*, *58*(10), 263-274.

254. Tokarz, M. (1990). On the logic of conscious belief. *Studia Logica*, *49*, 321-332.

255. Tokarz, M. (1993). *Elementy pragmatyki logicznej*. Warszawa: PWN.

256. Turri, J. (2010). Refutation by elimination. *Analysis*, *70*(1), 35-39.

257. Turri, J. (2011). The express knowledge account of assertion. *Australasian Journal of Philosophy*, *89*(1), 37-45.

258. Turri, J. (2012). Preempting paradox. *Logos & Episteme*, *3*(4), 659-662.

259. Turri, J. (2013). Knowledge guaranteed. *Noûs*, *47*(3), 602-612.

260. Turri, J. (2015a). Selfless assertions: some empirical evidence. *Synthese*, *192*, 1221-1233.

261. Turri, J. (2015b). Unreliable knowledge. *Philosophy and phenomenological research*, *90*(3), 529-545.

262. Unger, P. (1975). *Ignorance*. Oxford: Oxford University Press.

263. van Benthem, J. (2004). What one may come to know. *Analysis*, *64*(2), 95-105.

264. van Elswyk, P. (2023). Asking expresses a desire to know. *The Philosophical Quarterly*, pqad119.

265. van Elswyk, P., Benton, M. A. (2023). Assertion remains strong. *Philosophical Studies*, *180*(1), 27-50.

266. van Elswyk, P., Willard-Kyle, C. (forthcoming). Hedging and the Norm of Belief. *Australasian Journal of Philosophy*.

267. van Fraassen, B. C. (2020). Moore's Paradox Revenge. URL = <https://basvanfraassensblog.home.blog/2020/09/06/moores-paradox-revenge/>, accessed: 20.10.2022.

268. van Roojen, M. (2020). Promising and Assertion. In: Goldberg, S.C. (ed.) (2020). *The Oxford Handbook of Assertion*, Oxford: Oxford University Press.

269. Vanderschraaf, P., Sillari, G. (2022). Common Knowledge. In E. N. Zalta, U. Nodelman (eds.) *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition), URL = <https://plato.stanford.edu/archives/win2023/entries/common-knowledge/>.

270. Velleman, J. (1989). *Practical Reflection*, Princeton: Princeton University Press.

271. Von Wright, G. H. (1951). *An essay in modal logic*. Amsterdam: North Holland Publishing Company.

272. Wall, D. (2012). A Moorean paradox of desire. *Philosophical Explorations, 15*(1), 63-84.

273. Weatherson, B. (2004). Luminous Margins. *Australasian Journal of Philosophy*, *82*(3), 373-383.

274. Weiner, M. (2005). Must we know what we say?. *The Philosophical Review*, *114*(2), 227-251.

275. Willard-Kyle, C. (2020). Being in a Position to Know is the Norm of Assertion. *Pacific Philosophical Quarterly*, *101*(2), 328-352.

276. Williams, J. N. (1979). Moore's paradox: one or two?. *Analysis*, *39*(3), 141.

277. Williams, J. N. (1996), Moorean Absurdities and the Nature of Assertion, *Australasian Journal of Philosophy*, 74: 135–49.

278. Williams, J. N. (1998), Wittgensteinian Accounts of Moorean Absurdity, *Philosophical Studies*, 92: 283–306.

279. Williams, J. N. (2006). Wittgenstein, Moorean absurdity and its disappearance from speech. *Synthese*, *149*, 225-254.

280. Williams, J. N. (2007). Moore's paradoxes and iterated belief. *Journal of philosophical research*, *32*, 145-168.

281. Williams, J. N. (2013a). Moore's paradox and the priority of belief thesis. *Philosophical Studies*, *165*, 1117-1138.

282. Williams, J. N. (2013b). The completeness of the pragmatic solution to Moore's paradox in belief: a reply to Chan. *Synthese*, *190*(12), 2457-2476.

283. Williams, J. N. (2015). Eliminativism, Dialetheism and Moore's Paradox. *Theoria*, *81*(1), 27-47.

284. Williams, J. N. (2023). *A Unified Treatment of Moore's Paradox: Belief, Knowledge, Assertion and Rationality*. Oxford: Oxford University Press.

285. Williams, J. N., & Green, M. S. (2007). Introduction. In M. S. Green, J. N. Williams (Eds.), *Moore's Paradox: New essays on belief, rationality, and the first person*. Oxford: Clarendon Press.

286. Williamson, T. (1990). *Identity and discrimination*. Oxford: Blackwell.

287. Williamson, T. (1994). *Vagueness*. London: Routledge.

288. Williamson, T. (1996). Cognitive homelessness. *The Journal of Philosophy*, *93*(11), 554-573.

289. Williamson, T. (2000). *Knowledge and its limits*. Oxford: Oxford University Press.

290. Williamson, T. (2009). Replies to critics. In D. Pritchard, P. Greenough (eds.) (2009). *Williamson on Knowledge*. Oxford: Oxford University Press.

291. Williamson, T. (2013). Response to Cohen, Comesaña, Goodman, Nagel, and Weatherson on Gettier cases in epistemic logic. *Inquiry*, *56*(1), 77-96.

292. Williamson, T. (2014). Very improbable knowing. *Erkenntnis*, *79*, 971-999.

293. Williamson, T. (2017). Acting on knowledge. *Knowledge first: Approaches in epistemology and mind*, 163-181.

294. Williamson, T. (2020). Knowledge, credence, and the strength of belief. In: B. Reed, A. K. Floweree (Eds.) (forthcoming), *Towards an Expansive Epistemology: Norms, Action, and the Social Sphere*. London: Routledge.

295. Wittgenstein, L. (1998). *Remarks on the Philosophy of Psychology: Volume I*. G. E. M. Anscombe (trans.). Oxford: Blackwell.

296. Wittgenstein, L. (1999). *Philosophical Investigations: Second Edition*. G. E. M. Anscombe (trans.). Oxford: Blackwell.

297. Woods, J. (2014). Expressivism and Moore's Paradox. *Philosopher's Imprint, 14*(5), 1-12.

298. Woods, J. (2018). A commitment-theoretic account of Moore's paradox. In K. P. Turner, L. Horn (Eds.) *Pragmatics, Truth and Underspecification*. Leiden: Brill.

299. Yalcin, S. (2007). Epistemic modals. *Mind*, *116*(464), 983-1026.

300. Yalcin, S. (forthcoming). Defining common ground. *Linguistics and Philosophy*.

301. Yalcin, S. (unpublished manuscript). Common knowledge first.

302. Zakkou, J. (2018). The cancellability test for conversational implicatures. *Philosophy Compass*, *13*(12), e12552.

303. Zalta, E. N. (1988). *Intensional Logic and the Metaphysics of Intentionality*. Cambridge, MA: MIT Press.