Why the Causal Theory of Reference Fails to Immunize Metaphysical Realism Against Putnam's Model-Theoretic Arguments

Author: Pietro Lampronti

Degree: MSc Philosophy of Science

Supervisor: Dr. Wesley Wrigley

Year of Submission: 2024

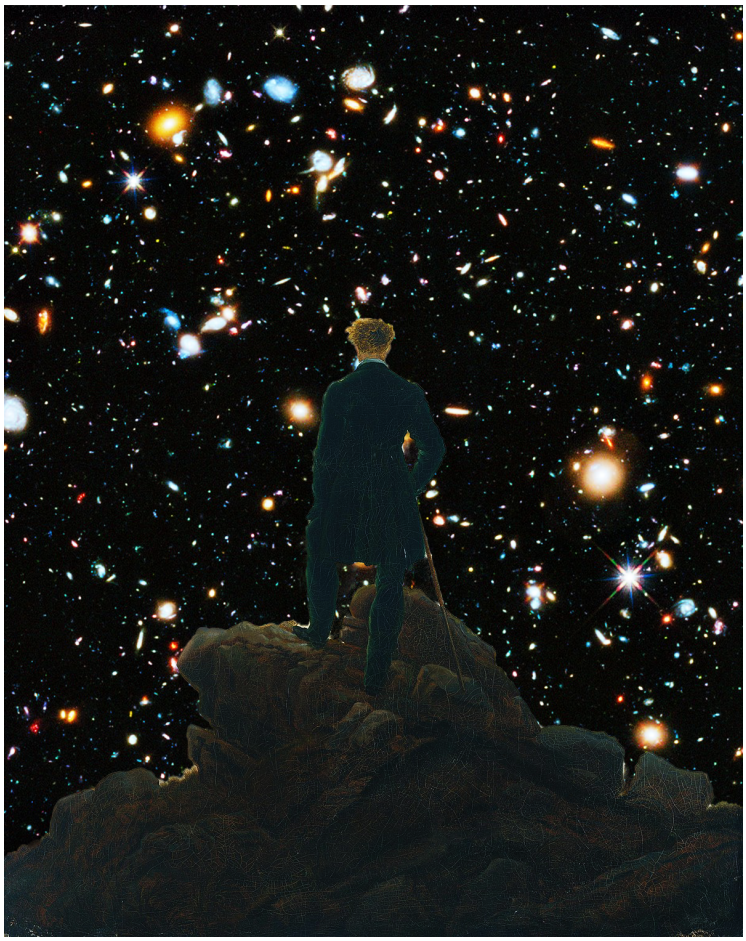
Word Count[1]: 10,000

[1]Including footnotes; excluding the table of contents and the bibliography.

"[...] *E quando miro in cielo arder le stelle;*
Dico fra me pensando:
A che tante facelle?
Che fa l'aria infinita, e quel profondo
Infinito Seren? che vuol dir questa
Solitudine immensa? ed io che sono? [...]"

"[...] *And when the stars' clear rays attract mine eyes,*
Within my soul I say:
What means so many a ray?
Where goes the wind? what booteth in the sky
The endless space serene? What is the thought
Of this vast solitude, and what am I? [...]"

Giacomo Leopardi,
Canto notturno di un pastore errante dell'Asia



Contents

1 Introduction

Starting in 1977, Putnam launched a series of model-theoretic attacks on a popular philosophical position he referred to as *Metaphysical Realism*. Such attacks were aimed at forcing (metaphysical) realists into a dilemma: either Metaphysical Realism is outright *incoherent*, since a logical contradiction can be generated from it, or it resorts to *magic*, and is thus unacceptable from a naturalistic perspective since it is devoid of empirical content. None of the two options is deemed acceptable by Putnam, who eventually comes to advocate for the replacement of Metaphysical Realism with his own brand of realism, internal realism.

The present work is aimed at evaluating whether the realist strategy of adopting the Causal Theory of Reference as a defence from Putnam's attacks is indeed successful in preserving metaphysical realism. Such a theory is the most popular among realists, for not only it is the best candidate among naturalistic ones, but it also is not an *ad hoc* defence, unlike other realist theories; for these reasons, demonstrating that it does not offer a viable way out of Putnam's dilemma would be a major blow to realists.

We will proceed as follows. First, we will outline Metaphysical Realism through its three tenets and give a model-theoretic account of it to facilitate its analysis. As it turns out, two questions are pivotal to the establishment of Putnam's dilemma and ultimately to the success or failure of his arguments: (a) whether language has a determinate reference relation with the world, or in model-theoretic terms, whether theories have an *intended* model (b) if this is the case, what fixes such a relation or model. A first attempt to answer both questions in favour of realists will be made through operational and theoretical constraints, which Putnam's model-theoretic attacks will however show to be insufficient to preserve Metaphysical Realism.

Then, we will illustrate the Causal Theory of Reference and how it promises to save Metaphysical Realism. Putnam's response to it, the infamous *just-more-theory* manœuvre, will be examined, which ventures to prove that the Causal Theory is as ineffective against his attacks as the aforementioned constraints and that invoking it amounts to begging the question. Amusingly, realists make the same accusation against Putnam and claim that he has distorted their view. This lays the grounds for the illustration of Putnam's dilemma: the Causal Theory is either (a) *empirical*, and is thus vulnerable to the *just-more-theory* manœuvre, ultimately acting as a basis to establish the incoherence of Metaphysical Realism, or (b) *magical*, and thus forces realists into an outlandish position that they are unwilling to hold.

Ultimately, for naturalistically minded philosophers, there seems to be no other option than that of embracing "the demise of a theory that lasted for over two thousand years" (Putnam; 1983: 74). What a time to be alive.

2 Metaphysical Realism

2.1 The Tenets of Metaphysical Realism

Metaphysical Realism ("MR") consists of three tenets: Independence, Correspondence, and Fallibilism.

Independence posits that there exists a world ("W") that is (mostly) made up of objects that are mind-, language-, and theory-independent (Button; 2013: 7-8) and that the world in turn (mostly) possesses those attributes too so that it is external, objective, and independent of us humans. The reason why we say "mostly" is that MR acknowledges that *some* objects, such as computers and cars, as well as thoughts, languages, etc., are artificial; however, they also hold that *most* objects, such as rocks, forests, stars, and galaxies, exist independently of us humans. In other words, Independence posits the existence of a "*ready-made* world" (Putnam; 1983: 211; emphasis in original): that is, it posits that W has an intrinsic, "built-in" structure or "furniture".

The second tenet of MR, Correspondence, claims that the concept of "truth" involves a "correspondence relation between words or thought-signs [of a language] and external things and sets of things [in W]" (Putnam; 1981: 49). Fleshing out a notion of MR that suits the purposes of the model-theoretic arguments, Correspondence claims that for a language $\mathscr{L}$ that contains names and predicates, there are correspondence relations that put names in $\mathscr{L}$ in correspondence with external things in W and predicates in $\mathscr{L}$ in correspondence with properties of and relations among things in W. Such relations can also be called interpretations or reference relations; the referent of a term according to a given relation is what that relation puts the term into correspondence with. By matching our thoughts and words with the constituents of W, reference relations enable us to think and talk about W.

Accordingly, we say that a sentence is true *relative to a given reference relation* if and only if what it says under that reference relation corresponds to what is the case in W. For instance, consider the reference relation R_1 that matches the term "snow" with snow and the predicates "white" and "black" with the properties of whiteness and blackness respectively. We would say R_1 is the standard reference relation that we normally use in everyday language—or at least that it is one of the ones that we use, the other ones perhaps differing in what they assign to terms other than the ones mentioned; let us adopt the working assumption that it is the only one that we use. Under R_1, the sentence "snow is white" is true, because it is the case that snow is white in W; that is, "snow is white" is true-*in*-R_1. Conversely, "snow is black" is false-*in*-R_1. Let us now consider the non-standard reference relation R_2 that matches "snow" with snow but inverts the meaning of the two predicates so that it matches "white" with the property of *blackness* and "black" with the property of *whiteness*. Under R_2, the truth values of the two sentences are inverted: "snow is white" is false-*in*-R_2 while "snow is black" is true-*in*-R_2.

Unsurprisingly, truth *relative to a given reference relation* depends on the given reference relation in question.

For truth *simpliciter*, the picture is different. Let us consider "snow is white"; as was pointed out earlier, what we normally want it to mean is that snow is white, and not that snow is black. That is, also recalling our working assumption, the *only* reference relation that we normally intend when we utter that sentence is R_1. In everyday talk, R_1 is the *intended* reference relation, as opposed to *unintended* ones, such as R_2; our working assumption thus says that the intended reference relation is *unique*, while there might be several *unintended* ones. More generally, the intended reference relation for a given sentence (in a given context, for a given speaker, etc.: call a sentence with such provisos "$sentence_P$". The provisos are important since the same sentence used in two different contexts might have two different intended reference relations) is whatever relation was intended for it, and the unintended ones are all the other ones; for instance, a non-standard speaker, or a standard speaker in a non-standard context (perhaps speaking in a code language with his friends), might utter "snow is black" with R_2 as the intended reference relation and R_1 as an unintended one. We can now define truth *simpliciter*: it is truth *relative to the unique intended reference relation*. Since in everyday talk, we adopt R_1 as the intended reference relation, in everyday talk "snow is white" is true *simpliciter* and "snow is black" is false *simpliciter*. Just as truth *relative to a given reference relation* depends on the given reference relation in question, truth *simpliciter* depends on the given *intended* reference relation in question.

Our working assumption that the intended reference relation, or interpretation, for a given $sentence_P$ is *unique* is one of the claims of realists (which, as we will see later, they have a hard time justifying). This makes the truth *simpliciter* of any $sentence_P$ non-relative, since it only depends on one interpretation. Now, let us define empirical theories of W as sets of $sentences_P$ about W: since each $sentence_P$ has a unique interpretation, so do the theories in question. MR leverages (a) this consequence of the working assumption and (b) Independence, which posits that W has an intrinsic furniture and structure, to claim that (c) there is a "One True Theory" of W (Button; 2013: 9), i.e., that there is just one theory of W that correctly describes W. That theory is the only one suited to be put, by its unique intended reference relation between its terms and W, into a complete correspondence with the intrinsic furniture and structure of W, such that it comes out true *simpliciter*. According to this claim, any other theory that is also put into a complete and true *simpliciter* correspondence with W by its unique intended interpretation is merely a "notational variant" of the One True one (Putnam; 1983: 211). We now see why MR requires W to have an intrinsic structure; if this was not so, then different theories could potentially correspond to W in different ways,[2] and truth might lose its non-perspectival nature (ibid.), *contra* the realist claim outlined earlier.

[2]From different perspectives; this view is called perspectival realism (Giere; 1994).

A consequence of Correspondence is that truth (of both the types mentioned) is non-epistemic (Putnam; 1977: 485), i.e., it is not contingent on whether someone knows, believes, or justifies it, but rather, it is a mind-independent relation between propositions, that are abstract and independent objects made of words or thought-signs, and external things and sets of things in W (van Inwagen; 1988: 104-107). In other words, something can be true even though it cannot be known to be so, and even what is most epistemically justifiable to believe might be false (Putnam; 1980: 473). For instance, one might utter the sentence S "there is snow in galaxy X-12" with R_1 as the intended interpretation, where "galaxy X-12" refers to some galaxy discovered by astronomers. Then, the truth *simpliciter* of S would not depend on whether anyone knew whether there was snow in galaxy X-12 or not; it would not even depend on whether someone in the history of humanity will ever know (maybe we will go extinct before discovering it).

To outline the third tenet of MR (again in a way that allows for the formulation of model-theoretic arguments), we first introduce a few notions. We earlier defined empirical theories of W as sets of $sentences_P$; to be more accurate, given a formal first-order language $\mathscr{L}$, we define an empirical *$\mathscr{L}$-theory* of W as a set of sentences couched in $\mathscr{L}$ that is closed under logical consequence that describes W. Since individual terms of $\mathscr{L}$ have several interpretations, so do $\mathscr{L}$-theories. One way to restrict the number of such interpretations is through *operational* and *theoretical* constraints. The former relate to the *empirical adequacy* of interpretations, i.e., the degree of coherence with the empirical data that is available: the number of empirical observations retrodicted and predicted and the degree of accuracy with which this is achieved. In other words, operational constraints restrict the set of interpretations to those that, when certain experiential conditions are satisfied, make certain sentences true *relative to the interpretations in question*. We say that (the interpretation of) an empirical theory fully satisfies operational constraints if and only if it is fully empirically adequate, that is, it retrodicts and predicts all empirical observations to the maximum degree of precision attainable by rational inquirers who experimented as much as possible. Theoretical constraints, on the other hand, relate to the formal properties of interpretations. Some of these constraints are logical consistency (i.e., whether an interpretation runs into contradictions); simplicity (e.g., of an interpretation to be understood or explained); elegance (e.g., how short and smooth the explanation an interpretation offers for certain linguistic or scientific phenomena is); explanatory power (i.e., how many linguistic or scientific phenomena an interpretation accounts for), etc. Since theoretical constraints oftentimes pull in different, if not opposite, directions, we imagine an equation in which each constraint has a coefficient and takes on a value, and say that (the interpretation of) an empirical theory fully satisfies theoretical constraints if and only if it maximises the overall weighted score of the equation. Finally, we say that (the interpretation of) an empirical theory of W is *ideal* if and only if it fully satisfies both operational and theoretical constraints.

We can now illustrate the third tenet of MR, Fallibilism, which claims that even an ideal empirical theory *might* be *false* (Douven; 1999: 479); more accurately, that even an ideal empirical theory *might* be false *simpliciter*, i.e., false according to its unique intended interpretation. Formulated in another way, the claim is equivalent to saying that for any empirical theory, being ideal is not sufficient for being the One True one. For instance, let us posit that W came into existence five minutes ago in such a way that we have false memories of a past that never existed, and let us further assume that we have come up with an ideal theory of W. It is perhaps reasonable to suppose that such a theory would satisfy, among other things, the operational constraint of retrodicting empirical observations of dinosaur fossils by positing the past existence of dinosaurs. If this was the case, we might have a strong belief that the theory was true *simpliciter*; but since truth is non-epistemic, it does not depend on what we believe, and the theory would still be false *simpliciter*.

Two important remarks on Fallibilism are due here. First, Fallibilism is a modal claim: it says that falsehood *simpliciter* ***might*** be possible, not that it is certain, or impossible. W might be as the unique interpretation of a given ideal empirical theory says it is, but it might also not. Second, Fallibilism is a claim about ideal *empirical* theories: the emphasis underlines the fact that it is *not* about theories that are non-natural, magical, nonscientific, or similarly characterized (Douven; 1999: 490).

For any ideal theory of W, if it truly had a *unique* intended interpretation, as MR states, it seems that Fallibilism would hold. To show this, suppose that we are in the foregoing skeptical scenario and that we are equipped with an ideal theory of W. Even if, similarly to the aforementioned snowy example, one was able to construct an *unintended* interpretation that made the theory true *relative to it* (perhaps by reinterpreting, among other things, talk of past dinosaurs in some other way), the ideal theory in question would still be false *simpliciter*, since the interpretation used to make it true *relative to it* would not have been the *intended* one. However, if for instance theories had *multiple* intended interpretations, the non-relative nature of truth *simpliciter* might be in danger. It is reasonable to argue that the definition of truth *simpliciter* according to MR would change from truth *relative to the intended interpretation* to truth *relative to the intended interpretation**s*** (plural); and different intended interpretations of a given theory might assign different truth values to parts, or even the whole, of it. It might even be the case that one of its intended interpretations makes it true *relative to it*, and so true *simpliciter*: if this was so, Fallibilism would be proved wrong.

As a matter of fact, we will prove that any ideal empirical theory of W is *guaranteed* to have several interpretations that make it true *relative to them*. Of course, it *might* happen that, by coincidence, one of such interpretations *is* the unique *intended* one; this is consistent with the modal nature of Fallibilism. However, if one of such interpretations were *guaranteed* to be the unique *intended* one, then Fallibilism would be refuted, for

it would be *guaranteed* that any ideal empirical theory would be true *simpliciter*. To preserve MR, realists need to argue that it *might* be the case that all such interpretations are *unintended*, i.e., that the unique *intended* interpretation *might not* be among such them.

2.2 A Model-Theoretic Account of Metaphysical Realism

Let us now give a model-theoretic account of MR, which will help in its analysis. To do so, let us introduce a few notions. Given a formal first-order language $\mathscr{L}$ and a set M of objects, we define the aforementioned correspondence/reference/interpretation relation as an *interpretation function* $\mathscr{I}$ that assigns (a) each name in $\mathscr{L}$ to an object in M and (b) each n-place predicate P in $\mathscr{L}$ to an extension, i.e., a subset of the n-fold Cartesian Product M^n. We then define an *$\mathscr{L}$-structure* $\mathscr{S}$ as an ordered pair consisting of (a) M as the domain set (b) the interpretation function $\mathscr{I}$. We say that a theory is *true in a structure* $\mathscr{S}$, or *true-in-$\mathscr{S}$*, if and only if every assertion that it makes about objects in M and properties and relations among them is correct. Finally, a *model* $\mathscr{M}$ of a theory is a structure that makes it true. For instance, let us define:

- The theory T: "The Earth is flat";[3]
- The structure $\mathscr{S}_1$ with (a) domain set {Earth} and (b) interpretation function assigning the name "Earth" to the Earth and the set {} to the predicate P "is flat";
- The structure $\mathscr{S}_2$ with (a) the same domain set as $\mathscr{S}_1$ and (b) interpretation function assigning "Earth" to the Earth and the set {Earth} to P.

In $\mathscr{S}_1$, the Earth does not belong to the extension of P; hence, "the Earth is flat" is false-in-$\mathscr{S}_1$. Conversely, since in $\mathscr{S}_2$, the Earth *does* belong to the extension of P, "the Earth is flat" is true-in-$\mathscr{S}_2$; $\mathscr{S}_2$ is thus a model of T.

Similarly to the criterion, outlined in the previous section, that establishes when a reference relation is to be called intended, when one has a theory intended to be made true by a certain structure (that is, by a certain domain set and a certain interpretation function), that structure is called the *intended* model of the theory (that domain set is the *intended* one, etc.). This is opposed to *unintended* models, i.e., other models of the theory that were however not its primary target. Again, similarly to the definition of truth *simpliciter* in terms of reference relations, in model-theoretic terms, truth *simpliciter* is truth *in the intended model* (in the *intended* domain set, etc.). For instance, let us define a third structure:

- $\mathscr{S}_3$ with (a) domain set {my frisbee} and (b) interpretation function assigning the name "Earth" to *my frisbee* and the set {my frisbee} to the predicate "is flat".

[3]While we acknowledge that this theory is not stated in a formal first-order language, we write as if it were for the sake of expository simplicity.

Let us posit that one uttered "the Earth is flat" with $\mathscr{S}_2$ as the intended model; since the Earth is indeed flat in $\mathscr{S}_2$, the theory would true-in-$\mathscr{S}_2$, i.e., true *simpliciter*. On the other hand, $\mathscr{S}_3$ would also be a model of T, albeit an unintended one. If another person knew that $\mathscr{S}_2$ was the intended model of T but did not know $\mathscr{S}_2$, and instead only knew $\mathscr{S}_3$, they could not pronounce themselves on the truth *simpliciter* of T; all they could say is that T is true-*in*-$\mathscr{S}_3$.

Now, starting from the external world W, we construct an $\mathscr{L}$*-structure* $\mathscr{W}$ consisting of (a) the set of external objects in W as domain set and (b) the unique intended reference relation between $\mathscr{L}$ and W as posited by MR as the interpretation function $\mathscr{I}$ assigning (i) each name in $\mathscr{L}$ to an object in W and (ii) a subset of W^n to each n-place predicate in $\mathscr{L}$. In other words, the unique intended reference relation $\mathscr{I}$ maps $\mathscr{L}$ onto W, such that each name and predicate of $\mathscr{L}$ has a determinate referent in W.

MR holds ideal theories, such as the One True Theory, to be theories of W; thus, the intended model of ideal theories is $\mathscr{W}$ (the intended domain set is that of $\mathscr{W}$ and the intended reference relation is $\mathscr{I}$). Fallibilism, which posits that even an ideal theory can be false *simpliciter*, can be reformulated as follows: even an ideal theory might be false-*in*-$\mathscr{W}$ (Melia; 1996: 172). So, if a realist were to evaluate the sentence "the Earth is flat" as part of the One True Theory, they would evaluate it with respect to the intended model $\mathscr{W}$; and since "the Earth is flat" is false-in-$\mathscr{W}$,[4] the realist would conclude that "the Earth is flat" is false *simpliciter*.

Now, recalling that "the Earth is flat" is true-in-$\mathscr{S}_2$ and true-in-$\mathscr{S}_3$, one legitimately ask: what makes $\mathscr{W}$ the *intended* model of the One True Theory but $\mathscr{S}_2$ and $\mathscr{S}_3$ (restrictions of) *unintended* ones?[5] To answer this question, realists would compare the differences between the three structures; to do so appropriately, they would restrict (a) its interpretation function to the terms that the interpretation functions of $\mathscr{S}_2$ and $\mathscr{S}_3$ also interpret, i.e., "Earth" and "is flat" and (b) the domain set of $\mathscr{W}$ to the set including only the thing that its interpretation functions assigns "Earth" to. Realists would claim that by doing so, they would obtain $\mathscr{S}_1$; in other words, they would claim that, for what concerns "Earth" and "is flat", the interpretation function of $\mathscr{W}$ operates in the same way as that of $\mathscr{S}_1$, thus assigning "Earth" to the Earth and "is flat" to the set {}. We can now compare $\mathscr{S}_1$, $\mathscr{S}_2$, and $\mathscr{S}_3$. First, we notice that the only difference between $\mathscr{S}_1$ and $\mathscr{S}_2$ is that they have a different interpretation function: that of the former assigns the set {} to "is flat", while that of the latter assigns "is flat" the set {Earth}. This is, realists would claim, what makes $\mathscr{S}_1$ (the restriction of) the *intended* model for the One True Theory and $\mathscr{S}_2$ (the restriction of) an *unintended* one: the former adopts the *intended* reference relation between $\mathscr{L}$ and W, while the latter adopts an *unintended* one. Second,

[4] At least according to the majority of scientists.

[5] It is clear enough why neither $\mathscr{S}_2$ nor $\mathscr{S}_3$ is not the intended model: because, for one thing, their domain set does not include all the things in W. However, one might wonder why one of such models is not *the restriction of* the intended model.

we notice two differences between $\mathscr{S}_1$ and $\mathscr{S}_3$: not only they have a different intended interpretation function, but they also have a different domain set. Realists would claim that $\mathscr{S}_3$ is (the restriction of) an *unintended* model because it has both an unintended domain set and an unintended interpretation function.

Now, by construction of $\mathscr{W}$, it is clear what its intended domain set is: the set of things in W. It is less clear what its intended *interpretation function* is: sure, it is the unique intended reference relation between $\mathscr{L}$ and W as posited by MR, but what is this reference relation anyway, and why is its restriction that of $\mathscr{S}_1$ and not that of $\mathscr{S}_2$ or $\mathscr{S}_3$? To answer this question, realists owe us an account of what determines such a reference relation; this will presumably also enable them to demonstrate that the intended reference relation is *unique*. Putnam's dilemma affirms that either there is no way to do so (MR is incoherent) or that all the ways to do so are implausible (MR is implausible).

3 What Fixes Reference?

If realists can carry out the onuses just outlined, they can safely preserve MR; as it turns out, they have a few options on the table. The rest of the work will examine three such options and the respective implications of their potential adoption.

3.1 Operational and Theoretical Constraints

It is best to start from the view that, for any ideal theory of W, it is operational and theoretical constraints that jointly single out its unique intended model $\mathscr{W}$, i.e., its intended domain set and its intended unique reference relation $\mathscr{I}$.

Putnam demonstrates this view to be false through the first part of his model-theoretic arguments. This part comes in two flavours: the *infallibilist* arguments, aimed at showing that any ideal theory has at least one model, and the *indeterminacy* arguments, which demonstrate that any theory that has at least one model has several models (Button; 2013). As we will show, the difference between such models can lie in either the domain set or the interpretation function. While any difference in domain sets can be used relatively straight-forwardly to single out the *intended* model among *unintended* ones, we will see that it is differences in the interpretation functions that turn out to be problematic.

Let us start with a first infallibilist argument, and assume we select, out of the set of ideal theories of W that are formalizable in First-Order Logic with Identity ("$FOL^{=}$"), the theory T_1.[6] The realist would claim that T_1, despite being ideal, might still be false

[6]The assumption that such a set is non-empty, potentially including all ideal theories, might be worth a more in-depth analysis. Nonetheless, most commentators raise no concern (e.g., Taylor (1991: 152)). As Douven (1999: 488) points out in a brief footnote, Putnam argues that even if ideal theories require higher-order logics to be formalized, the model-theoretic arguments still work, albeit in a different form. Presumably, the one scenario in which they would not work is if ideal theories were not formalizable at all; but such a scenario does not seem plausible.

simpliciter, i.e., false-in-$\mathscr{W}$. Since T_1 is ideal, and being logically consistent is one of the theoretical constraints required for being ideal, T_1 is consistent. By the Strong Completeness Theorem of $FOL^{=}$,[7] T_1 has a model $\mathscr{M}$: T_1 is true-in-$\mathscr{M}$. Since the choice of T_1 was arbitrary among ideal theories, we have shown that any ideal theory has at least one model.

At this point, realists might object to the nature of the entities in the domain set of $\mathscr{M}$ (Taylor; 1991: 153). As a matter of fact, due to how the proof of the Strong Completeness Theorem works, $\mathscr{M}$ is a mere "abstract synthetic construction" (Resnik: 1987: 151): that is, it is a *numerical* model, its domain set consisting entirely of natural numbers. Such a set differs quite significantly from the intended domain set of $\mathscr{W}$, which consists of the things in W. So, realists might claim that (a) since $\mathscr{M}$ has an *unintended* domain set, it is merely an *unintended* model of T_1 (b) as a direct consequence, T_1 being true-in-$\mathscr{M}$ does neither mean nor imply that T_1 is true-in-$\mathscr{W}$, i.e., true *simpliciter* and (c) Putnam has not shown that T_1 is true-in-$\mathscr{W}$.

Unfortunately for the realist, the same line of reasoning does not work on the second infallibilist argument, which goes as follows. First, we assume that W contains infinitely many things (for instance, it can be broken down into infinitely many regions of space-time), and let the size of that infinite be the cardinality κ of W. Now, Resnik (1987: 151) points out how an elementary theorem of model theory that he calls "hidden inflation theorem" can be used, if necessary, to increase the size of $\mathscr{M}$ until a new model $\mathscr{M}_1$ is obtained that has cardinality κ, i.e., the same as W.[8] $\mathscr{M}_1$ can now be mapped onto W to generate a new model $\mathscr{M}_2$ of T_1. This is done by defining a bijection between $\mathscr{M}_1$ and W, i.e., a bijection between the objects of $\mathscr{M}_1$ and those of W and between the relations among objects of $\mathscr{M}_1$ and those among objects of W.[9] By construction, the domain set of $\mathscr{M}_2$ consists of the things in W, and the bijection between $\mathscr{M}_1$ and W is the interpretation function of $\mathscr{M}_2$, i.e., the reference relation between $\mathscr{M}_2$ and W: we call this relation, the *satisfaction relation SAT*.

By construction, all the structures mentioned so far in this section are models of T_1: T_1 is true-in-$\mathscr{M}$, true-in-$\mathscr{M}_1$, and true-in-$\mathscr{M}_2$. So, for any ideal theory, we have shown that it is possible to generate several models of it. Some of them will clearly be *unintended* since they have *unintended* domain sets (such as $\mathscr{M}$ and $\mathscr{M}_1$), while some other, as we will see shortly, *might* be the *intended* one. At this stage, what is important to notice is that all such models can be generated while leaving the truth-conditions of sentences of the theory in question completely unchanged. In other words, while operational and theoretical constraints do determine the truth of sentences of theories, they cannot determinately fix the *reference* of the terms in such sentences: "*truth-conditions* for *whole sentences*

[7] Strong Completeness Theorem of $FOL^{=}$: any consistent first-order theory has a model (Button, 2013: 16).

[8] See also Taylor (1991: 153). An alternative way to obtain $\mathscr{M}_1$ is illustrated by Putnam (1977: 485).

[9] For technical details, see Button (2013: Appendix I).

underdetermine reference" (Putnam; 1980: 35; emphasis in original).

Now, let us compare $\mathscr{M}_2$ with $\mathscr{W}$. By construction, they have the same domain set, i.e., the set of things in W: $\mathscr{M}_2$ thus has the *intended* domain set for models of ideal theories. Furthermore, they both have a reference relation with W: by construction, $\mathscr{M}_2$ refers to W through SAT, while $\mathscr{W}$ does so through the unique intended interpretation relation $\mathscr{I}$. The question then is: are $\mathscr{M}_2$ and $\mathscr{W}$ the same model, i.e., *is* $\mathscr{M}_2$ $\mathscr{W}$? Since they share the domain set, the only part in which the two models can differ is in the interpretation function, i.e., the reference relation with W. So, the question above can be reformulated as: is SAT $\mathscr{I}$?

Let us consider the issue from the perspective of truth. $\mathscr{M}_2$, similarly to $\mathscr{M}$, is a model of T_1; thus, T_1 is true-in-$\mathscr{M}_2$. So, if $\mathscr{M}_2$ is $\mathscr{W}$, then truth-*in*-$\mathscr{M}_2$ is truth-*in*-$\mathscr{W}$; then T_1 is true-in-$\mathscr{W}$. An equivalent way to put it is the following: since $\mathscr{M}_2$ is a model of T_1 and the reference relation between $\mathscr{M}_2$ and W is SAT, $\mathscr{M}_2$ makes T_1 "true" of W, provided that we interpret this notion of truth as "$true_{SAT}$"; true-in-$\mathscr{M}_2$ and $true_{SAT}$ are then equivalent formulation. So far, the two infallibilist arguments show that there is a way to make any ideal theory $true_{SAT}$. This is potentially threatening for Fallibilism, which claims that any ideal theory might be false *simpliciter*. Since truth *simpliciter* is truth *in the intended model*, which for ideal theories such as T_1 is $\mathscr{W}$, if SAT turns out to be $\mathscr{I}$, $\mathscr{M}_2$ turns out to be $\mathscr{W}$, and so truth *simpliciter* turns out to be $truth_{SAT}$; if this is the case, then the two infallibilist arguments show there is a way to make any ideal theory true *simpliciter*, refuting Fallibilism.

So far, realists agree with Putnam; specifically, they agree that for any ideal theory, it is guaranteed that there is at least one model with the *intended* domain set (in this case, $\mathscr{M}_2$) that makes the theory true *relative to it*. Of course, they concede, it might even happen that, by mere coincidence, the reference relation of the model in question is also *intended* (SAT might be $\mathscr{I}$); in this case, such a model would effectively be the *intended* one ($\mathscr{M}_2$ would be $\mathscr{W}$), and the ideal theory would be true *simpliciter*. What realists deny is that this is not guaranteed to be the case: that is, the reference relation of the model in question *might not* be the *intended* one. This allows them to maintain the claim that such a model might be an *unintended* one, and thus that the ideal theory in question, despite being true *relative to the model in question*, might still be false *simpliciter*, thus ultimately upholding Fallibilism.

To sustain the assertion that SAT might not be $\mathscr{I}$, however, realists owe us an account of what determines reference relations. This task is of the same type as the one which we opened this section with: explaining why it is (the restriction of) the intended reference relation of $\mathscr{S}_1$ that is the *intended* one and not that of $\mathscr{S}_2$ or $\mathscr{S}_3$. We conclude that operational and theoretical constraints indeed fail to determine reference uniquely: realists must find another answer.

3.2 The Causal Theory of Reference

A potentially promising route to explain how reference is determinately fixed is to posit that while operational and theoretical constraints determine truth-conditions for sentences of an ideal theory but leave their reference underdetermined, it is *extra-linguistic* constraints that determine the reference of such sentences (Brueckner; 1984: 137). In Putnam's words, this amounts to saying that reference is fixed by "non-psychological" constraints, i.e., by "nature itself" (Putnam; 1983: xii). The best candidate among such constraints is the Causal Constraint ("CC"), which states that there are causal connections between the language of an ideal theory and W (Button; 2013: 20) that only the *unique intended* reference relation satisfies. Such connections are those posited by the Causal Theory of Reference ("CTR"), first advanced by Kripke (1980), which holds that *causation (uniquely) fixes reference*: in other words, the terms of a language acquire determinate referents through an initial "naming ceremony", i.e., an initial act of naming, and maintain their determinate referents since each successive use of such terms is causally chained to the initial ceremony. For instance, the first moment in which newborns are referred to by a given name is their naming ceremony (*for that name*; they might have been previously referred to by expressions such as "my incoming baby", etc.), and causal chains unfolding thereafter that link any subsequent use of the name in question to the initial naming ceremony preserve its reference.

Let us illustrate the functioning of CC through our foregoing example of T_1. We concluded the section on operational and theoretical constraints by asking ourselves whether $\mathscr{M}_2$ and $\mathscr{W}$ were identical, i.e., whether SAT and $\mathscr{I}$ are. Now, CC posits that for ideal theories, $\mathscr{I}$ is the only reference relation satisfying the required causal connections. If SAT also does, it means that it is the same reference relation as $\mathscr{I}$, so $\mathscr{M}_2$ is $\mathscr{W}$ and T_1 is true *simpliciter*. As mentioned above, this possibility is accepted by realists, who instead maintain that SAT *might not* be $\mathscr{I}$, since, due to its abstract model-theoretic construction, it *might not* satisfy the required causal connections. In this case, it would be an *unintended* reference relation; $\mathscr{M}_2$ would be an *unintended* model, T_1 might still be false *simpliciter*, and Fallibilism would be preserved. In other words, if SAT and $\mathscr{I}$ were different, CC would allow us to restrict the class of models of T_1 as determined by operational and theoretical constraints to just its *unique intended* model ($\mathscr{W}$) with its *unique intended* reference relation ($\mathscr{I}$).

There are several reasons why CC is the best candidate among extra-linguistic constraints. The first one relates to what Douven (1999: 479) dubs the principle of *Semantic Naturalism*, which states that semantics is as *empirical* as any other science; therefore, so are theories of reference, which then are part of *empirical* ideal theories too (as we will see in the rest of the work, this conclusion is of the uttermost importance for the model-theoretic arguments). The principle is considered by Putnam to be a minimum

requirement for the substantiveness and plausibility of a theory of reference and thus for the constraint on reference built on it, for a theory that did not satisfy it would be of "dubious ontological status" (Anderson; 1993: 315). If such a theory were accepted, it would simply be the case that ontological implausibility would replace semantic heterogeneity; but in Putnam's view, the former is as unacceptable as the latter (ibid.). Requiring theories of reference to be empirical amounts to requiring that reference "must ultimately supervene upon a *naturalistic* relation" (ibid.: 313; emphasis mine), as several physicalists have upheld: the mechanism that fixes reference must therefore be definable in naturalistic terms. The reason why CC is the best candidate among *naturalistic* extra-linguistic constraints is that causation is the best, if not the only, candidate mechanism for reference fixing satisfying Semantic Naturalism (Douven; 1999: 487). Second, unlike others, such as the Eligibility Constraint proposed by Lewis (1984) (which posits that there are more and less *eligible* referents for the terms of an ideal theory), CC relies on a theory of reference, CTR, that has been advanced independently of Putnam's arguments; therefore, its underpinning is general and not *ad hoc*. Third, it applies regardless of whether the assumption that ideal theories can be formalized in $FOL^{=}$ holds or not, unlike constraints that verge on technical details regarding second-order logic or modal logic and the notion of possible worlds (Button; 2013: 23).

Unsurprisingly, CC is the most popular constraint adopted by realists:[10] because of this, demonstrating that it does not suffice to determine reference would be a major blow for MR.

3.3 The *Just-More-Theory* Manœuvre

Of course, Putnam attempts to do just that, through the so-called *just-more-theory* ("JMT") manœuvre.[11]

As we have seen, realists agree with Putnam up to the point that operational and theoretical constraints underdetermine reference; that is, it is agreed by both that reference has not yet shown to be determinedly fixed. At this stage, realists mention CC and maintain that it does the trick. Putnam then considers if doing so, i.e., *saying* CC, fixes reference; in other words, he considers whether the linguistic formulation of CC guarantees not only its own referential determinacy (Anderson; 1993: 315), but referential determinacy overall. JMT ventures to prove that this is not the case. As a matter of fact, if realists think that it is causation that fixes reference, i.e., that CC is true, they will have to incorporate it into their ideal theory T_1, [12] because Semantic Naturalism requires

[10] E.g., Devitt (1983) and Brueckner (1984).

[11] To be clear, JMT is leveraged by Putnam against other types of constraints as well, except the Magical Constraint (Button; 2013: 21-23). We will see why such a constraint is exempt from JMT later.

[12] Provided that CC is formalized in $FOL^{=}$ in a way that is logically consistent in itself and with T_1. Just like the assumption any ideal theory can be formalized in $FOL^{=}$, that the same can be done with CC is not clear (Douven; 1999: 488; footnote 21). According to Resnik (1987: 155), Putnam might claim,

theories of reference to be empirical and so part of ideal theories. By doing so, they will have obtained the new ideal theory $T_2 = T_1 \cup \{CC\}$. At this point, the same procedure used earlier to generate unintended models of T_1 can be applied to T_2; this will result in the generation of a reference relation SAT_2 and in the guarantee that T_2 is $true_{SAT_2}$. This result is problematic for realists, since just as they maintained that SAT *might not* be $\mathscr{I}$, they want to maintain that SAT_2 might not be $\mathscr{I}$. Another way to put this is the following: one might generate several *unintended* models of T_2, and in each model $\mathscr{N}$, $reference_{\mathscr{N}}$ is fixed by $causation_{\mathscr{N}}$; however, unless the word "causation" already has a determinate reference to a relation across all models—and the question of how this is possible for any term is the very question at stake— it will refer to a different relation in each model, therefore failing to determine a determinate extension for "reference" across all models (Putnam, 1980: 477). Including CC in T_1, i.e., *saying* CC, is thus shown to be ineffective in fixing reference.

The issue above is nicely conveyed through a parable (Putnam; 1983: ix-xi), here adapted to a thought experiment. Imagine there are two different models $\mathscr{O}_1$ and $\mathscr{O}_2$ of ordinary English such that men adopt $\mathscr{O}_1$ and women adopt $\mathscr{O}_2$, and that both models satisfy both operational and theoretical constraints, such that they adopt the same truth assignment to English sentences; in particular, both models assign truth to "causation fixes reference". Suppose the difference in the two models lies not in their domain set but in their interpretation function so that the terms "causation", "fix", and "refer" get interpreted differently, and that two philosophers, a man and a woman, both claim that "causation fixes reference". We note that while the man would mean that $causation_1$ $fixes_1$ $reference_1$, the woman would mean that $causation_2$ $fixes_2$ $reference_2$. Because $\mathscr{O}_1$ and $\mathscr{O}_2$ adopt the same truth-assignment to English sentences, it would be impossible for the two philosophers to realize that they would be meaning different things and be talking past each other. Moreover, they would both be right, each in their own way, since $causation_1$ $fixes_1$ $reference_1$ is true-in-$\mathscr{O}_1$ and $causation_2$ $fixes_2$ $reference_2$ is true-in-$\mathscr{O}_2$. Thus, unless the linguistic formulation of CC already has determinate reference across all models, *saying* CC will not fix reference.

For these reasons, Putnam states that realists who affirm that it is causation *itself* that fixes reference are *assuming* that there exists some "safe conceptual heaven" (Taylor; 1991: 161) into which they can retire to safely formulate CC. Such realists are giving for granted that, when they say "causation", they are *determinately* referring to the "real", *unique* causation: they are thus *assuming* that their words have a *unique intended* interpretation to *prove* so, ultimately advancing a circular argument and begging the question. In other words, realists are "ignoring [their] own epistemological position" (Putnam; 1983: xi).

By the same model-theoretic procedure that guarantees the existence of an interpreta-

and has possibly done so, that a constraint that is not so formalizable is not "genuine"; however, no proof of this is offered in his presentations of the model-theoretic arguments.

tion that makes any ideal theory true *relative to it*, we have shown that the existence of an interpretation that makes CC true *relative to it* is also guaranteed, because if CC is to be upheld it is to be part of ideal empirical theories. This point is similar to that with which we ended sections 2.1 and 3.1. Two other realist considerations will be similar. First, just as with ideal empirical theories, it *might* happen that, by coincidence, the foregoing interpretation of CC *is* the unique *intended* one. This leads us to the second consideration: despite advancing operational and theoretical constraints first and the causal one second, it seems that realists have not yet proved that, just as with whole ideal empirical theories, the interpretation that makes CC true *relative to it* whose existence is guaranteed *might not* be the CC's unique *intended* one. If they want to preserve MR, this is what realists need to conclusively show now.

3.4 The Realist Response

Arguably, the best way to do so is to defend CC by refuting JMT. Lewis (1984) attempts to do exactly so: he operates a distinction between "satisfying C[C]-theory [and] conforming to C[C]" (1984: 225), and claims that JMT confounds the former with the latter.[13] He agrees with Putnam that the linguistic formulation of CC guarantees neither its own referential determinacy nor referential determinacy overall: since it is embedded in T_2, the linguistic formulation of CC acts "from *within*" the theory (Van Cleve; 1992: 349), and it is evident that it is as liable to misinterpretations as the rest of the theory. What Lewis disagrees with is that, when formulating CC, realists mean that it is such a linguistic formulation that fixes reference; instead, what they posit is that it is causation *itself* that fixes reference. The role of the linguistic formulation of CC is merely to inform us of this, and not to fix reference itself. CC is not merely a linguistic formulation but a *extra-linguistic* constraint: as such, it acts "from *outside*" T_1 (ibid.), and is thus not liable to the misinterpretations of the theory.

Lewis attributes the cause of Putnam's misunderstanding to the dialectic of philosophy, which generally favours the sceptic. In this context, Putnam, the reference sceptic, asks realists what determinately fixes reference; realists reply by saying CC. The sceptic then claims that, because of the model-theoretic arguments, unless the linguistic formulation of CC already had determinate reference, it still has no determinate reference, since the mere act of uttering it does not make it acquire determinate reference; so he once again challenges realists to demonstrate that reference is determinate, specifically concerning the linguistic formulation of CC. Realists are checkmated: for they can neither repeat CC nor mention another constraint, on pain of receiving the same response from the sceptic and be dragged into a hopeless infinite regress. As Devitt (1983: 298) points out, no matter what answer is given to a certain question, the sceptic will always be able to ask another

[13] Resnik (1986: 156-157), Anderson (1993: 314), and Douven (1999: 485-486) argue along similar lines.

question about the meaning of the answer just received; nonetheless, "explanation must stop somewhere" (ibid.), and the fact that an answer can be challenged by the sceptic does not in itself imply that it was not a legitimate answer to the question posed (ibid.: 299). As a matter of fact, realists might have been right the first time: if causation indeed fixed reference, CC would indeed have had determinate reference, and it would also have been true. Unfortunately, even if realists were right, there would be no way for them to win the argument against the sceptic.

Devitt rightly notes that "we cannot say anything about the relationship between language and the world without saying something, i.e., without using language" (ibid.: 298): there is no way for the realist to express CC but formulate it in the words of their language, since language is the only option we have to express thoughts (Taylor; 1991: 165). Of course, such words will have several interpretations, as JMT shows; but if the constraint they express is true, then their reference will be determinate, and all the interpretations just mentioned but one—their *intended* one—will not conform to the constraint and will thus be *unintended*. In this case, there would be facts of the matter as to the determinacy of reference regardless of whether we had formulated a theory of reference or not, and if so, regardless of whether it were true or not (Heller; 1988: 123). Putnam would ask what fixes the reference of "causation" (and of "fixes", and of "reference"); the realist will reply that, if CC is true, it is *causation itself* that fixes the reference of "causation". Determinate reference is possible thanks to causal relations, and not to "metaphysical glue". Devitt argues that Putnam is liable of *assuming* that CC is *wrong* instead of *showing* so: this, in turn, leads to the conclusion that the reference of its linguistic formulation is not determinate, which prompts the sceptical question of what fixes its reference. Because of this, Devitt concludes that Putnam begs the question against the realist; for to prove the point that the linguistic formulation of CC lacks determinate reference, he is not authorized to assume that CC is false (ibid.; 299).[14]

3.5 Is The Causal Theory of Reference Magic in Disguise?

Let us now momentarily cast aside CC, for the time has come to introduce the only extra-linguistic constraint that *Putnam* claims would allow the realist to refute his arguments: the Magical Constraint "MC". Just as CC states that there are causal connections between the language of an ideal theory and W that only the unique intended interpretation of such a language satisfies, MC states that there are *magical, intrinsic* connections between... (Button; 2013: 30). Such connections are those posited by a *Magical* Theory of Reference ("MTR"). For instance, one might posit that it is "noetic rays" that determine magical connections and thus the reference of terms in the language (Putnam; 1981: 51). Another example is extreme Platonism, which posits that we have a non-natural—and to

[14]See also Devitt (1984: 276).

Putnam, mysterious (1980: 466)—mental power that allows us to directly "grasp" forms or "concepts" (ibid.: 464).

The reason why MC is successful in refuting Putnam's arguments is that it leverages a theory of reference, MTR, which is *magical* and not *empirical*. This places it outside the scope of Fallibilism, which claims that any ideal *empirical* theory might be false *simpliciter*; thus, realists who uphold MTR need not be committed to the view that such a theory *might* be false *simpliciter* (Douven; 1999: 490), and can safely claim that it *is true simpliciter*. This allows them to safely maintain that the intended interpretation of ideal *empirical* theories is *unique* and that it *might not* be among the ones that make such theories true *relative to them* whose existence is guaranteed by the model-theoretic arguments, ultimately preserving Fallibilism.

However, after entertaining MC, Putnam discards it. It turns out that the strength of MC is also its weakness: for its being *magical* allows it, on the one hand, to escape Fallibilism, but on the other hand, it prevents it from satisfying the principle of Semantic Naturalism, which claims that theories of reference are to be *empirical*. Upholding MC thus means replacing semantic heterogeneity with the equally unacceptable option of ontological implausibility. To naturalistically minded philosophers, MC and MTR are pieces of useless from an epistemological perspective and unconvincing from a scientific one (Putnam, 1980: 471). Positing that a non-naturalistic mechanism such as noetic rays determines reference amounts to relapsing into "medieval essentialism" (1983: xii) by positing that "a *one-knows-not-what* [...] solves our problem *one-knows-not-how*" (ibid.; emphasis in original). He deems the epistemological problems generated by such a view as obstacles that cannot be surmounted (1981: 16) and concludes that no contemporary philosopher would dare to hold it (ibid.: 51).

The reason why we introduced MC is that Putnam thinks CC is just another type of MC (Devitt; 1984: 275), and challenges realists to show otherwise. As Anderson (1993: 321) points out, "for too long philosophers have gotten away with a wink and a nod parading as a substantive theory of reference": Putnam's challenge to realists consists in elaborating on such a wink and such a nod to demonstrate that CTR is indeed substantive, where this is understood as having empirical content, i.e., satisfying Semantic Naturalism, and thus being fundamentally different from MTR. This, of course, while avoiding falling into the trap of JMT, which attracts all empirical theories. As a matter of fact, Putnam sets up realists in a dilemma: they can either (a) formulate a theory of reference that is indeed *empirical*, and thus satisfies Semantic Naturalism, but precisely because of Semantic Naturalism, it must be included in ideal theories, and is thus vulnerable to JMT, or (b) formulate a *magical* theory of reference to avoid the strangle of the model-theoretic arguments, however failing to satisfy Semantic Naturalism and therefore the very realist standards for what constitutes acceptable semantics.

To be sure, the linguistic formulations of CC and MC are indeed different: the latter

posits *magical* connections, while the former posits *causal* ones. However, so far, "causation" has acted as a mere placeholder; but coming up with a name for a concept that is then left undefined does not in itself represent a substantial explanatory advance (Anderson; 1993: 316). For all we have been told, "causation" might be a concept as magical as "noetic rays" or "non-natural grasp". To show that it is not, realists need to further explain just *how* causation fixes reference, and to do so in such a way that allows them to establish that CC, unlike MC, has empirical content.

To further illustrate the dilemma, let us consider just *how* causation is supposed to fix reference. Realists typically claim that while no word *necessarily* corresponds to one thing rather than another, it is *contextual*, and therefore contingent, causal connections between a word and a thing that fix the correspondence between them in the unique intended interpretation of the language of a speaker. It is implicitly assumed that the thing to which a given word refers is the dominant cause of beliefs about that thing that contain such a word. Here Putnam (1951: 51) raises the first issue: he considers how the dominant cause of belief in electrons is textbooks, and yet that the word "electron" does *not* refer to textbooks, but to *electrons*. If this were so, the foregoing implicit assumption would reveal itself to be false. To avoid this undesirable conclusion, realists claim that it is causal chains *of the appropriate type* that determinately fix reference; and since the word "electron" is supposedly *not* connected to textbooks by one of such chains, it is perfectly normal that it does not refer to textbooks, even if the latter are the dominant cause of belief containing it.

At this point, the obstacle realists need to clear is to give an account of what counts as a causal chain *of the appropriate type*; and this turns out to be quite the challenge (Button; 2013: 21). Several potential definitions, both Humean and non-Humean, are discussed by Putnam in his writings,[15], but all of them are found to be defective. Not so much because they are defective, but because there are several tentative definitions of causation, Anderson (1993)—on Douven's (1999: 486) reconstruction of his argument—argues that since causation is overdetermined, it cannot act as a reference fixer.

Even granted that the problem above was solved and that a widely agreed upon naturalistic definition of causation was provided, realists still need to explain how the reference of words that refer to things we have no causal connection (of the appropriate type) with is fixed. According to Putnam, there are two different types of such words. The first one is words which refer to things that we have no causal connection with but that are *of the same kind* as other things that we do have causal connections with. For instance, when we use the term "tiger", we also refer to tigers that we have no causal connection with, such as *future* tigers, i.e., tigers that will exist in the future. According to

[15] A non-comprehensive list includes: leaving causation as a primitive, as suggested by Boyd (1980); defining causation non-physicalistically; Mill's concept of *total cause*; and Lewis' (1973) definition through counterfactuals. Putnam examines these in (1983; 211-217).

realists, what enables us to refer to future tigers is that they are *of the same kind* as past and current ones—namely, all of them are tigers. The first tenet of MR, Independence, comes into play here: for the notion of an object being 'of the same kind' as another only makes sense within a categorial system which says what counts as a similarity (Putnam; 1981: 53), and for such a similarity, and thus for reference, to be determinate, such a system must be intrinsic to W. The second type of words that refer to things we have no causal connection with that we are nonetheless able to refer to is that of words whose kind we currently do not have, never had, and never will have causal connections with, such as the word "extraterrestrial".[16] In this case, the realist response consists of two steps. The first is to renounce the claim that there are causal connections between *everything* we are able to refer to and us, limiting the existence of such connections to *basic* terms and their referents. The second step is to uphold that such basic terms can be combined together to obtain descriptions of non-basic terms that refer to things we have no causal connections with. In the case of "extraterrestrial", realists would claim that we have causal connections with *terrestrials*, the property "sentient being", and the relation "from another planet", and that these collectively suffice to build a complex description that introduces and fixes the reference of "extraterrestrial". Of course, this explanation requires in turn an account of what counts as a *basic* term. However, this lies beyond the scope of the present work; so does the explanation of how we are able to refer to *non*-physical things, and things that have no causal role (Button; 2013: 21), such as mathematical objects and aesthetic values—as a hint, Putnam (1983: 205) claims that providing such an explanation usually proves to be a difficult challenge.

To wrap up, this elucidation of CTR affirms the existence of three ways in which we are able to refer: (a) by causal connections of the appropriate type (b) through the fact that some objects which we do not have any causal connections with are of the same kind as objects we do have causal connections of the appropriate kind with (c) by description consisting of basic terms whose referents we have causal connections of the appropriate type with. Moreover, as pointed out, some obstacles encountered along the way are yet to be cleared. So: what to make of CTR? Putnam argues that it seems "not so much false as otiose" (1981: 53). In other words, he thinks that it lacks empirical content and that it is therefore "*empirically indistinguishable*" from, and thus equivalent to, an MTR (Button; 2013: 31).

Another path that leads to the same conclusion is offered by the examination of the version of CTR proposed by Field (1972), who posits that reference is a determinate and unique "physicalistic relation, i.e., a complex causal relation between words or mental representations and objects or sets of objects", which is to be discovered by empirical science (Putnam; 1981; 45). At first, such an account surely seems to fulfil Semantic Naturalism quite satisfactorily, if anything thanks to all the naturalistic keywords present

[16] See footnote 3.

in it. Let us call the posited physicalistic relation R and suppose that the following is true:

(1) *x refers to y* if and only if *x bears R to y*

Putnam calls Field's view into question by asking what *makes* (1) true. He rightly points out that there are, as previously demonstrated, several reference relations between words and things, and wonders what *singles out* R as the determinate and unique one. Having found no satisfying answer, he proceeds to claim that "the fact that R *is* reference [seems to] be a *metaphysically unexplainable* fact, a kind of primitive, surd, metaphysical truth" (ibid.: 46; emphasis in original), and that "believing that some correspondence intrinsically just *is* reference [...] amounts to a[n MTR]" (ibid.: 47; emphasis in original).

Ultimately, as Douven (1999; 486) points out, Putnam's relegation of CTR to the realm of magic does not so much turn on technical deficiencies or difficulties that might beset such a theory as on the structural impossibility of conquering immunity against JMT all the while satisfying Semantic Naturalism. On the other hand, we explored what happens in the other horn of the dilemma earlier in the work: if an *empirical* account of CTR is given, CTR *is* indeed just more theory, and realists have run out of options to formulate a reference-fixing constraint. Through the model-theoretic procedures deployed against ideal empirical theories, several interpretations of CC can be generated, some of which make it true *relative to them* and some of it make it false *relative to them* (Douven; 1999: 482): truth then becomes "radically indeterminate" (Brueckner; 1984: 135). To be fair, not only *truth* would be indeterminate, but the *reference* of *all* words would be too; that is, radical referential indeterminacy would be obtained. As Anderson (1993: 313) points out, Putnam does not even remotely entertain the possibility of embracing referential indeterminacy: he considers the intrinsic incoherence of such a view as an argument in itself against it.

If realists were unable to explain how *determinate* reference is possible without resorting to either just more empirical theory or magic, they would be unable to explain what determines the unique *intended* interpretation of any ideal empirical theory too (Brueckner; 1984: 136). This, in turn, would deprive them of the grounds on which they would be entitled to claim that such an interpretation *might not* be among the "model-theoretic" ones, ultimately forcing them to abandon this view. On these grounds, any ideal empirical theory is *guaranteed* to come out true *in some* ***intended*** *interpretation of it*, and so true *simpliciter*, contradicting Fallibilism, which states that even an ideal empirical theory *might* be false *simpliciter*. Ultimately, MR is proved "incoherent" (Putnam; 1977: 483) and "collapse[s] into *unintelligibility*" (ibid.; 486; emphasis in original).

4 Conclusion

The time has now come to summarize our adventurous journey through the meanders of Metaphysical Realism. Its formulation in model-theoretic terms has acted as a basis for Putnam's model-theoretic arguments to be deployed, which demonstrate that any ideal empirical theory is *guaranteed* to have some interpretations that make it true *relative to them*. Realists have been challenged to preserve Fallibilism, i.e., to demonstrate that it *might* be the case that none of them is the *intended* interpretation of such a theory, and so that the theory might still fall short of truth *simpliciter*. While having failed to do so through operational and theoretical constraints, we have witnessed how realists attempt to achieve their goal through the promising Causal Theory of Reference.

Things now become intriguing. Mutual accusations of begging the question are raised by both parties: Putnam does so through the *just-more-theory* manœuvre, while realists adopt the strategy of claiming to have been misinterpreted. At this point, the *Magical* Theory of Reference has been introduced, followed closely by Putnam's challenge to the realists to demonstrate that it is *not* equivalent to the Causal one. Ultimately, this lays the ground for the establishment of Putnam's dilemma: either the Causal Theory is *empirical*, which allows, through an elaborate series of steps, to prove that Metaphysical Realism is *incoherent*, or it is *magical*, which forces realists to renounce their beloved principle of semantic naturalism, and adopt a theory of reference that is bereft of empirical content.

Ultimately, Putnam does *not* prove that Metaphysical Realism is *false* (Anderson: 1993: 321). What he proves is that if it is to be spared from outright incoherence, it must incur into "medieval essentialism". This, at least to naturalistically minded philosophers, is perhaps as untenable a position as incoherence; such philosophers are thus forced to abandon Metaphysical Realism. Putnam's *reductio ad absurdum* is complete (ibid.: 313).

We have thus witnessed "the demise of a theory that lasted for over two thousand years" (Putnam; 1983: 74). Well; as another, perhaps equally illustrious philosopher once wrote, *the times they are a-changin'*.

Bibliography

Anderson, D. L. (1993) "What Is the Model-Theoretic Argument?". The Journal of Philosophy. [Online] 90 (6), 311–322.

Block, N. (1980) "Readings in the Philosophy of Psychology". Cambridge, Mass.

Boyd, R. (1980) "Materialism Without Reductionism: What Physicalism Does Not Entail". In Block, N. (1980), 67-106.

Brueckner, A. L. (1984) "Putnam's Model-Theoretic Argument Against Metaphysical Realism". Analysis (Oxford). [Online] 44 (3), 134–140.

Button, T. (2013) "The Limits of Realism". 1st ed. Oxford: Oxford University Press.

Devitt, M. (1983) "Realism and the Renegade Putnam: A Critical Study of Meaning and the Moral Sciences". Noûs (Bloomington, Indiana). [Online] 17 (2), 291–301.

—— (1984) "Reviewed Work(s): Reason, Truth and History by Hilary Putnam". The Philosophical Review. 93 (2), 274-277.

Douven, I. (1999) "Putnam's Model-Theoretic Argument Reconstructed". The Journal of Philosophy. [Online] 96 (9), 479–490.

Field, H. (1972) "Tarski's Theory of Truth". The Journal of Philosophy. [Online] 69 (13), 347–375.

French, P. et al. (1988). "Realism and Antirealism". Midwest Studies in Philosophy 12. Minneapolis: University of Minnesota Press.

Heller, M. (1988) "Putnam, Reference, and Realism". Midwest Studies in Philosophy. [Online] 12 (1), 113–127.

Kripke, S. A. (1980). "Naming and Necessity: Lectures Given to the Princeton University Philosophy Colloquium". Cambridge, MA: Harvard University Press. Edited by Darragh Byrne & Max Kölbel.

Lewis, D. (1973). "Counterfactuals". Oxford: Blackwell.

—— (1984) "Putnam's Paradox". Australasian Journal of Philosophy. 62 (3), 221 – 236.

Melia, J. (1996). "Against Taylor's Putnam". Australasian Journal of Philosophy. 74 (1), 171 – 174.

Putnam, H. (1977) "Realism and Reason". Proceedings and Addresses of the American Philosophical Association. [Online] 50 (6), 483–498.

Putnam, H. (1980) “Models and Reality”. The Journal of Symbolic Logic. [Online] 45 (3), 464–482.

—— (1981) “Reason, Truth, and History”. Cambridge [Cambridgeshire]; Cambridge University Press.

—— (1983) “Realism and Reason”. Cambridge: Cambridge University Press.

Resnik, M. D. (1987) “You Can’t Trust an Ideal Theory to Tell the Truth”. Philosophical Studies. [Online] 52 (2), 151–160.

Taylor, B. (1991) “‘Just More Theory’: A Manoeuvre in Putnam’s Model-Theoretic Argument for Antirealism”. Australasian Journal of Philosophy. 69 (2), 152 – 166.

Van Inwagen, P. (1988) “On Always Being Wrong”. In French et al. (1988), 95-111.