METAPHYSICS

Simchen, O. 2019. Realism and Instrumentalism in Philosophical Explanation. *Metaphysics*, 2(1), pp. 1–15. DOI: https://doi.org/10.5334/met.20

RESEARCH

Realism and Instrumentalism in Philosophical Explanation

Ori Simchen
University of British Columbia, CA
ori.simchen@ubc.ca

There is a salient contrast in how theoretical representations are regarded. Some are regarded as revealing the nature of what they represent, as in familiar cases of theoretical identification in physical chemistry where water is represented as hydrogen hydroxide and gold is represented as the element with atomic number 79. Other theoretical representations are regarded as serving other explanatory aims without being taken individually to reveal the nature of what they represent, as in the representation of gold as a standard for pre-20th century monetary systems in economics or the representation of the meaning of an English sentence as a function from possible worlds to truth values in truth-conditional semantics. Call the first attitude towards a theoretical representation *realist* and the second attitude *instrumentalist*. Philosophical explanation purports to reveal the nature of whatever falls within its purview, so it would appear that a realist attitude towards its representations is a natural default. I offer reasons for skepticism about such default realism that emerge from attending to several case studies of philosophical explanation and drawing a general metaphilosophical moral from the foregoing discussion.

Keywords: realism; instrumentalism; philosophical explanation; philosophical methodology

1 Introduction

Any theoretical endeavour employs representations and philosophy is no exception. A representation for present purposes is a theoretical apparatus that stands for some subject matter within a purported explanation. So a representation in the relevant sense is before all else an explanatory tool. This is obviously very rough but it helps delineate my topic and distinguish it from neighbouring topics in the philosophy of mind and language. 'Representation' as used here applies to a plethora of explanatory devices; 'represent' to a plethora of explanatory activities. The models of a particular model theory, for example, are representations in the relevant sense to the extent that they are employed in the course of explaining matters of interpretability for formal theories. A formal language for such a theory is itself a representation in the relevant sense in its facility to explain syntactic properties and relations for a fragment of natural language. And the model theory for such a formal language, representing as it does word-world relations and such emergent properties and relations as truth and entailment, is a representation in the intended sense as well. Given such heterogeneity, we do well to settle for the rough characterization. For present purposes we stay clear of the general matter of what confers representationality on representations.[1] Our rough characterization is broadly functional.

We regard theoretical representations either as revealing the nature of the represented themselves, or else as representing the represented for other theoretical purposes without such revelatory pretension. For lack of better terminology I call the first way of regarding a representation *realist* and the second way *instrumentalist*. A clear case of the former kind is the way we regard the representation of gold as a transition metal within physical chemistry. The substance is thus represented under a widespread realist attitude

[1] For my own take on such questions, see my (2017).

towards the representation: being a transition metal is part of what it is to *be* gold.[2] It figures prominently in our account of the nature of the substance. A case of the latter kind is the way we regard the representation of gold as a standard for pre-20th century monetary systems within economics. The substance is thus represented theoretically under a widespread instrumentalist attitude: being a standard for pre-20th century monetary systems is not part of what it is to be gold but is nevertheless instrumental to the explanation of gold's economic significance. Even if being such a standard plays an important role in revealing something important about the nature of the economy, it is not presumed to reveal something important about the nature of the represented substance individually. In contrast to the way the terms 'realist' and 'instrumentalist' often characterize substantive theses in the metaphysics of truth, and in contrast to the way they might characterize attitudes towards whole theories, the use of this terminology here is with respect to attitudes towards individual theoretical representations, however they are ultimately individuated, within a broader theoretical context.

The question of realist vs. instrumentalist attitude can be raised for representations wielded within various philosophical explanations as well. Philosophical explanation is typically taken to reveal the nature of whatever falls within its purview, so it would seem that a realist attitude towards its representations is a natural default. My main aim in this paper is to offer reasons for skepticism about such default realism, reasons that emerge from attending to several case studies – numbers, *de re* modality, cognitive attitudes – and drawing a general metaphilosophical lesson from the foregoing discussion. The emphasis is on the qualifier 'default' and the focus is, once again, on individual theoretical representations. The upshot is not meant to be that philosophical explanation is not after all in the business of uncovering the nature of things; it is rather that many familiar representations wielded within philosophical explanation should not be taken individually to reveal the nature of whatever they represent.

The question of realist vs. instrumentalist attitude towards theoretical representations demands that we regard representations apart from what they represent. There is a special difficulty distinguishing the representation and the represented within various philosophical explanations, a difficulty bred by the abstractness of the various subject matters of philosophical reflection *inter alia*. Much of the paper – Sections 2–4 – will be taken up by prying apart representations from represented in several areas of philosophical concern, illustrating realist vs. instrumentalist choice points along the way. As will emerge in Section 5, the representations deployed within our case studies all have a distinctly logico-semantic origin. The significance of this will be brought out in the ensuing discussion of whether a realist attitude toward theoretical representations is warranted. I will argue that in none of the cases discussed is a realist attitude warranted despite the clear presence of realist purport. This has broader implications for philosophical explanation that will be taken up in the concluding Section 6.

A word of caution before we proceed. The history of modern philosophy is replete with debates surrounding realism. If history is to be our guide, there are good reasons to remain pessimistic when it comes to whether the question of realism can be resolved effectively via some *a priori* considerations.[3] It seems unlikely that some powerful new such consideration is in the offing to resolve the matter once and for all. A promising way out is to change the terms of the discussion surrounding realism. First, rather than asking directly whether a realist thesis is true for a given domain, we consider whether a realist attitude towards a bit of theory concerning the domain is warranted, keeping in mind cases for which there is widespread *de facto* agreement. Second, and relatedly, cases for which there is widespread *de facto* agreement on a realist attitude provide a useful point of comparison when it comes to whether such an attitude is warranted in more controversial cases. Third, and finally, in keeping with a general avoidance of *a priori* considerations pertaining to the broad question of realism, the main issue of whether a realist attitude towards theoretical representations in philosophy is warranted is approached piecemeal via case studies. The present inquiry casts doubts on whether such representations should be taken in a realistic spirit by default. This is achieved by examining specific cases of philosophical explanation that are both central enough to the overall enterprise and yet distant enough from one another per topic to be usefully representative.

It is, of course, understood that the present strategy will seem of limited reach from a more traditional mindset on the question of realism, especially given that the ur-concern of truth or falsity for the thesis of realism is not directly addressed by the approach. For example, there is widespread agreement on a realist attitude when it comes to the representation of water as hydrogen hydroxide within physical chemistry. A

[2] It is also part of what it is to be palladium, part of what it is to be iridium, part of what it is to be silver, etc.

[3] A prominent attempt at such resolution is found in Putnam's work. See especially Putnam (1977), (1978), and (1981).

traditionalist might insist that this fact by itself is inconsequential when it comes to the question whether water *really does* have the particular molecular nature alleged by physical chemistry. From the present point of view, however, the fact that being hydrogen hydroxide is revelatory of the true nature of water is on a surer footing than the doubts of the traditionalist on the question of realism. We do well to ask after characteristics present in a theoretical identification such as water being hydrogen hydroxide and use them as a measure for explanations proffered in philosophy. This is what I propose to do here. As we are about to see, the procedure can yield interesting results.

2 Frege on Number

Let us turn to the three case studies of philosophical explanation that will occupy us throughout, beginning with what Dummett (1993) characterizes as the original site of the linguistic turn in philosophy: Frege's ingenious construction of the natural numbers within his second-order logic as equivalence classes of first-level concepts under the relation of equinumerosity.[4] In a nutshell, Fregean concepts are the referents of predicates. In other words, concepts are semantic values, a point to which we will return in due course. A first-level concept is a function from objects to truth-values, a second-level concept is a function from first-level concepts to truth-values, and so on. Equinumerosity is a second-level relational concept mapping first-level concepts ϕ and ψ to the True just in case there is a bijection between the ϕs and the ψs, otherwise mapping them to the False.[5] A number n is construed as the extension of the second-level concept *equinumerous with* $\underline{\phi}$, where ϕ is a first-level concept with an n-membered extension.[6] Such is Frege's basic construction. We now consider two distinct attitudes towards it.

First, in keeping with Frege's original attitude we might say, ambitiously, that the extension of the second-level concept *equinumerous with* $\underline{x \neq x}$ (call it **0**) is what the number zero turns out to be, the extension of *equinumerous with* $\underline{x = \mathbf{0}}$ (call it **1**) is what the number one turns out to be, the extension of *equinumerous with* $\underline{x = \mathbf{0} \vee x = \mathbf{1}}$ (call it **2**) is what the number two turns out to be, and so on. On this approach, the extension of any of these second-level concepts reveals what the relevant natural number really is at bottom.[7] As the original proposal faces well-known problems, it is consonant with the present attitude to find fixes that would salvage as much of the original idea as possible and maintain that the patched up version does a better job at identifying what the numbers really are. Once the fixes are found, we can say what sort of things numbers are, much like water being hydrogen hydroxide or gold being the element with atomic number 79. So maintains the proponent of a realist attitude towards Frege's original construction.

There can be little doubt that Frege's own attitude towards his construction was realist in our sense. This may initially seem difficult to maintain in light of some things Frege says about definitions. In *Grundlagen* we read:

> The definition of an object does not, as such, really assert anything about the object, but only lays down the meaning of a symbol. After this has been done, the definition transforms itself into a judgment, which does assert about the object; but now it no longer introduces the object, it is exactly on a level with other assertions made about it. (Frege 1953: §67)

This would appear to make Frege's definitions of the numbers ill-suited to reveal their natures. The definition of the number zero as the extension of the second-level concept *equinumerous with* $\underline{x \neq x}$, for example, is officially offered as laying down the meaning of the numeral '0', laying bare the sense of that sign that presumably was previously implicit. Once this has been achieved, Frege claims, the judgment that zero is the extension of *equinumerous with* $\underline{x \neq x}$ becomes just another truth (albeit a trivial one) about zero, on par with other judgments, such as that zero immediately precedes one in the number series. This does not sound like someone who is concerned to tell us what zero really is.

---

[4] The construction is found in Frege's (1953) *Grundlagen*. Later, in *Grundgesetze*, Frege (2013) defines the numbers as equivalence classes of classes under equinumerosity, but this change has little direct bearing on the present discussion.

[5] So just in case $\exists\xi(\forall x(\phi x \rightarrow \exists!y(\psi y \wedge \xi xy)) \wedge \forall x(\psi x \rightarrow \exists!y(\phi y \wedge \xi yx)))$, where $\exists!vFv$ abbreviates $\exists v(Fv \wedge \forall u(Fu \rightarrow u = v))$.

[6] I use italics for mentioning concepts – a difficult issue broached in Frege's (1952) and contemplated ever since – and underlining for the mentioning of concepts within the mentioning of concepts (double concept-quoting, as it were). I mostly set aside well-known problems with logicist reductions, such as the notorious contradiction arising from Frege's Basic Law V or the violations of the Axiom of Foundation engendered by the variant proposal that identifies the number n with the class of n-membered classes, even though I do discuss briefly the neologicist reaction to the aforementioned contradiction towards the end of this section.

[7] For a contrasting reading of Frege that emphasizes a more pragmatic strain in his thinking, see Reck (2007). My present focus is on the more common reading of Frege's project as endorsed by much of the philosophical tradition it spawned.

But how is the judgment that zero is the extension of the second-level concept *equinumerous with* $\underline{x \neq x}$ really on par with other judgments about the number? Consider the fact that zero immediately precedes one in the number series. This fact is for Frege the fact that for some first-level concept ϕ and some object y, ϕ is in the extension of the second-level concept *equinumerous with* $\underline{x = \mathbf{0}}$ while the first-level concept ϕ*-other-than-*y is in the extension of the second-level concept *equinumerous with* $\underline{x \neq x}$. This depends on the number zero being identified with the extension of *equinumerous with* $\underline{x \neq x}$. Once we attend to such matters of dependence it is difficult to avoid the conclusion that being the extension of *equinumerous with* $\underline{x \neq x}$ reveals the nature of the number itself – it is what enables the obtaining of the posterior arithmetical facts concerning zero, including the fact that zero immediately precedes one in the number series. Indeed, such facts of dependence make it implausible that Frege could think of his explicit definitions of the individual numbers as anything but revelatory of the nature of the numbers themselves. We can see this more clearly by delving into some of the details.

The definition Frege (1953: §76) offers of the relation of immediate succession in the number series, $n = S(m)$, is:

(S) $$\exists \phi \exists y \left(\phi y \wedge \hat{X}(\approx_x (Xx, \phi x)) = n \wedge \hat{X}\left(\approx_x (Xx, \phi x \wedge x \neq y)\right) = m \right).$$

(In words: for some first-level concept ϕ and some object y, n is the number belonging to ϕ and m the number belonging to ϕ*-other-than-*y.) From this definition, together with the explicit definition of zero as the extension of *equinumerous with* $\underline{x \neq x}$, i.e. $\hat{X}(\approx_x (Xx, x \neq x))$, and of one as the extension of *equinumerous with* $\underline{x = \mathbf{0}}$, i.e. $\hat{X}(\approx x (Xx, x = \mathbf{0}))$, it easily follows that $\mathbf{1} = S(\mathbf{0})$. For Frege (1953: §73) proves what is now known as Hume's Principle,

(HP) $$\hat{X}\left(\approx_x (Xx, \phi x)\right) = \hat{X}\left(\approx_x (Xx, \psi x)\right) \leftrightarrow \approx_x (\phi x, \psi x).$$

(In words: the extension of *equinumerous with* $\underline{\phi}$ is the extension of *equinumerous with* $\underline{\Psi}$ just in case ϕ and ψ are equinumerous.) He then proves that $\approx_x (x = \mathbf{0} \wedge x \neq \mathbf{0}, x \neq x)$. From the definitional $\hat{X}(\approx_x (Xx, x \neq x)) = \mathbf{0}$ and (HP) it then follows that $\hat{X}(\approx_x (Xx, x = \mathbf{0} \wedge \mathrm{x} \neq \mathbf{0})) = \mathbf{0}$. Taking ϕ as $x = \mathbf{0}$ and y as $\mathbf{0}$, the relevant instance of (S) follows immediately from the definitional $\hat{X}(\approx_x (Xx, x = \mathbf{0})) = \mathbf{1}$:

$$\exists \phi \exists y \left(\phi y \wedge \hat{X}\left(\approx_x (Xx, \phi x)\right) = \mathbf{1} \wedge \hat{X}\left(\approx_x (Xx, \phi x \wedge x \neq y)\right) = \mathbf{0} \right).$$

The logicist claim that zero immediately preceding one in the number series is a logical truth thus dovetails the logicist construction of zero and one as the logical objects **0** and **1**. It is difficult to imagine Frege's insistence on the logicality of zero immediately preceding one in the number series while demurring at the further suggestion that $\hat{X}(\approx_x (Xx, x \neq \mathrm{x}))$ and $\hat{X}(\approx_x (Xx, x = \mathbf{0}))$ reveal what zero and one really are. Frege's attitude towards these constructions seems clearly realist in our sense.[8] Turning a familiar Quinean quip on its head, we see here that essence (of the number zero, say) is what meaning (of the numeral '0') becomes when it is divorced from the word and wedded to the object of reference.[9]

Putting Frege's realist predilections to one side, we might say more cautiously and out of step with Frege's attitude that **0** represents the number zero, that **1** represents the number one, that **2** represents the number two, etc. – all as part of an overall effort to show that arithmetic need not avail itself of any mathematical means beyond second-order logic. On such an approach, the extensions of the second-level concepts represent the natural numbers for a broader explanatory purpose. But like the representation of gold as the standard for pre-20th century monetary systems in economics, they are not themselves presumed to reveal the nature of the numbers.

Each of these attitudes towards Frege's construction, the realist and the instrumentalist, has something going for it. We need not rehearse the benefits of realism here – the grand aspirations of logicism speak for themselves and have had a momentous impact on subsequent philosophy. But the instrumentalist attitude can, while the realist attitude cannot, straightforwardly accommodate firm intuitive verdicts that both lay people and working mathematicians pass on the numbers, routine verdicts that are at odds with basic features of Frege's construction taken realistically. For example, when we ascribe the number seventeen to a

[8] For a more Carnapian reading of Frege's construction, see Reck (2007). It should be clear that I disagree with Reck's reading.

[9] "Meaning is what essence becomes when it is divorced from the object of reference and wedded to the word" (Quine 1951: 22).

collection of things we think of the collection as in some sense *having* this numerical attribute, of seventeen *belonging* to the collection, or being in some sense *in* the collection as a whole. But on the Fregean construal taken realistically it would be more apt to say that the collection – or rather the characteristic function of the set associated with it, the first-level concept – belongs to **17** rather than the other way around (in a different sense of 'belong', of course). According to the instrumentalist attitude, then, **17** represents the number seventeen but shouldn't be expected to reveal the nature of what it represents. That the characteristic function of the set of seventeen things "belongs" to **17** is an "artifact of the model" in Kaplan's sense.[10] In short, Frege's **17** is one thing, the number seventeen is another – all in good Butlerian fashion.[11]

In the foregoing discussion we ignored the issue of paradox that famously besets Frege's logicism. We might, however, view the neologicist evasion of paradox as itself an instrumentalist wrinkle in Frege's overall logicist project.[12] The neologicist discards Frege's Basic Law V that permits the derivation of unrestricted comprehension and ultimately leads to contradiction. The fundamental move is to take the following version of Hume's Principle as an axiom schema instead, where '$\#(\phi x)$' abbreviates 'the number of ϕs':

$$(\mathrm{HP}^{\mathrm{neo}}) \qquad \#(\phi x) = \#(\psi x) \leftrightarrow \approx_x (\phi x, \psi x).$$

This claim that the number of ϕs is the number of ψs just in case ϕ and ψ are equinumerous is thought to give a contextual definition of the meaning of the cardinality operator '#' ("the concept of Number"). With suitable adjustments to Frege's definitions we can then prove the axioms of arithmetic without any reliance on Basic Law V. We can prove $1 = S(0)$, for example, if we define immediate succession along Fregean lines as:

$$(S^{\mathrm{neo}}) \qquad \exists\phi\exists y\left(\#(\phi x) = n \wedge \#(\phi x \wedge x \neq y) = m\right).$$

We then proceed exactly as before except under the representation of zero and one as $\#(x \neq x)$ and $\#(x = 0)$, respectively. For example, the proof in Frege (1953: §75), which can be utilized for the intermediary claim that $\approx_x (x = 0 \wedge x \neq 0,\ x \neq x)$, nowhere relies on the explicit definitions of the numbers.[13] But the neologicist's $\#(x \neq x)$ and $\#(x = 0)$ need not be thought of as revealing the natures of zero and one, respectively. We may regard them as representing zero and one for the broader explanatory purpose of demonstrating the logicality of such facts as that $1 = S(0)$. In other words, we can take an instrumentalist attitude towards such individual representations of the numbers. This easily coheres with the neologicist program.

3 *De Re* Modality

A second illustration of the choice between realist and instrumentalist attitude towards theoretical representations in philosophy emerges in the context of debates in the 1960s and 1970s surrounding the metaphysics of modality. Earlier on, in the 1950s and 1960s, logicians were developing model theories for systems of first-order quantification plus non-extensional operators. The basic insight in the logic of modality emerged from early observations that logically speaking 'possibly' and 'necessarily' are duals, much like 'some' and 'all'. Just as $\exists$ is definable as $\neg\forall\neg$ or $\forall$ as $\neg\exists\neg$, the diamond of possibility and the box of necessity are inter-definable: $\Diamond$ as $\neg\Box\neg$ or $\Box$ as $\neg\Diamond\neg$. This structural commonality suggested that $\Diamond$ and $\Box$ are specialized quantifiers. An index set was provided for them to range over: a set of "possible worlds". $\Diamond p$ is true just in case for some possible world w, p is true at w. $\Box p$ is true just in case for every possible world w, p is true at w. The importance of this idea for subsequent metaphysics of modality is difficult to exaggerate. Much of the contemporary metaphysical engagement with possibility, necessity, and cognate notions (impossibility, contingency, actuality, etc.) is unfathomable without the precedent set by this basic semantic idea.

---

[10] Kaplan (1975) writes:

> When we construct a model of something, we must distinguish those features of the model which represent features of that which we model, from those features which are intrinsic to the model and play no representational role. The latter are artifacts of the model. For example, if we use string to make a model of a polygon, the shape of the model represents a feature of the polygon, and the size of the model may or may not represent a feature of the polygon, but the thickness and three-dimensionality of the string is certainly an artifact of the model. (722)

[11] "Everything is what it is, and not another thing" – Bishop Butler.

[12] See Heck (2011) for a comprehensive up-to-date assessment of the current state of neologicism, a project inspired by Geach (1955) and initially carried out in further detail by Parsons (1965) and Wright (1983).

[13] The proof is that, given that no object falls under $x = 0 \wedge x \neq 0$, and given that no object falls under $x \neq x$, every relational concept is such that every object falling under the one is so related to a unique object falling under the other, and vice versa. So *some* relational concept is such. And so, the concepts are equinumerous.

It is not feasible to offer a comprehensive survey of the various ramifications of the possible worlds semantic apparatus for the metaphysics of modality. We focus instead on a single aspect: the quantified modal logical capture of *de re* modality. Sidestepping technical details, the syntactic contrast between the constructions $\Diamond\exists x\phi x$ and $\exists x\Diamond\phi x$ standardly receives the following treatment in the possible world semantics for quantified modal logic: $\Diamond\exists x\phi x$ is true just in case for some possible world w, some individual in the domain of w is a member of the extension of ϕ in w; $\exists x\Diamond\phi x$ is true just in case for some actual individual and some possible world w, the individual in question is a member of the extension of ϕ in w. The former construal may or may not require for its truth a non-actual individual ϕ-ing in some w.[14] But our present focus is on the latter: what does it mean to say of some actual individual, say an individual who happens not to ϕ, that for some possible world *w that very individual* is a member of the extension of ϕ in w? What does the identification of an individual in another possible world amount to? Take Kripke's (1980) example of losing the 1968 US presidential election. Someone, Nixon, say, lacks the property – Nixon actually won the 1968 election – but might have had it – the 1968 election was a close one. So we have $\exists x(\neg\phi x\wedge\Diamond\phi x)$ and the entailed $\exists x\Diamond\phi x$ is made true by Nixon, the actual winner, losing in a counterfactual situation. But how can this be? Nixon actually lacks ϕ but has ϕ in the counterfactual situation, we are assuming, so by the indiscernibility of identicals the actual Nixon is distinct from the counterfactual Nixon. And yet did we not just say that Nixon himself, despite not ϕ-ing, might have ϕ-ed? Thus goes the problem of transworld identity.[15]

Let us consider two different general attitudes towards the possible world representation of the *de re* modal fact that Nixon might have lost, a realist attitude and an instrumentalist attitude – call them *A* and *B*, respectively. Attitude *A* takes the possible world construal of the *de re* modal fact that Nixon might have lost the 1968 election to reveal the nature of the modal fact in question.[16] For there to be a possible world where Nixon loses is what it is for Nixon to possibly lose. According to this attitude, the problem of transworld identity poses a genuine metaphysical perplexity that demands an answer. If inhabiting a world where he loses is what Nixon's possible loss really is, at bottom, then we need to explain how his apparent transworld existence does not violate the indiscernibility of identicals. Here there are two broad strategies to consider.

The first emanates from the observation that the problem arises from the monadic character of properties the having of which needs to be relativized to worlds. On this first approach, while the property of losing the 1968 election is monadic when considered on its own, for Nixon to have the property in the counterfactual situation is for Nixon to bear the dyadic relation *x*-losing-in-*w* to the world in question. Correlatively, for Nixon to lack the property in actuality is for Nixon to fail to bear this relation to the actual world. That Nixon has the property in *w* and lacks it in actuality no longer speaks to the distinctness of the counterfactual Nixon and the actual Nixon, any more than my being at rest relative to the floor and in motion relative to the earth's axis speaks to my self-distinctness. And here, once again, one can take two distinct attitudes towards the proposal, call them subsidiary attitudes A_1 and A_2.

According to the realist subsidiary attitude A_1, for Nixon to bear the dyadic relation to the counterfactual situation is just what it is for Nixon to have the property of losing in that situation. This is what having a monadic property in a world turns out to be upon closer theoretical scrutiny – a dyadic relation to a world. But according to the instrumentalist subsidiary attitude A_2, Nixon bearing or failing to bear the *x*-losing-in-*w* relation to a world merely represents what it is for the property of losing to be itself monadic when considered on its own while for something to have it or not to have it is relativized to a world.[17] According to this second attitude A_2, the dyadic relation is not really what the instantiation of the monadic property of losing in *w* amounts to – it merely represents monadic instantiation in *w* for the broader theoretical purpose of capturing *de re* modal predication.[18]

[14] Much ink has been spilled over cases where ϕ seems neither instantiated by anything actual nor possibly instantiated by anything actual while $\Diamond\exists x\phi x$ seems true. I set this aside here, but see my (2013) for my considered take on the issue.

[15] A further development of the issue is the early epistemological concern with identifying the individual in the counterfactual situation, which we set aside.

[16] Such an attitude is clearly exhibited in what Plantinga (1976) calls "the canonical conception of possible worlds".

[17] One might, for example, suppose on independent grounds that the relata of relations in *re* are particulars and deny that possible worlds are such.

[18] The situation under A_2 may thus seem quite different from the situation regarding my being in motion relative to the earth's axis but at rest relative to the floor. In the latter case it is often supposed, realistically, that the properties of being in motion or being at rest, while apparently monadic, are revealed upon closer theoretical scrutiny to be relations borne to reference frames. Whether or not in this case there is wiggle room for preserving the monadic character of being in motion or being at rest while maintaining relativity as per the formalism of SR is an interesting question that cannot be pursued here. See, however, some preliminary discussion in Section 5 concerning theoretical identifications in natural science.

Now, if we ignore the instrumentalist option encapsulated in A_2 and focus only on realist subsidiary attitude A_1 we might recoil from the present suggestion once the monadic properties under consideration seem sufficiently intrinsic. This gives rise to a second strategy for meeting the problem of transworld identity. Consider Lewis's (1986) example of a five-fingered hand being possibly six-fingered. The present approach under attitude A_1 would render the unactualized possibility of being six-fingered for the actually five-fingered hand to turn out to be a matter of the hand being related to a counterfactual situation. But being six-fingered seems intrinsic enough that it can seem perniciously revisionary to maintain that for the hand to have this property in a possible world is for the hand to bear a relation to the world in question. Having such a property is a matter intrinsic to the thing having it.

Famously, the Lewisian recoil from the present approach – considered again under attitude A_1 – is to maintain that for the actually five-fingered hand to be possibly six-fingered is for the hand to have a six-fingered counterpart in another possible world. By extension, for Nixon to possibly lose the 1968 election despite actually winning is for Nixon to have a counterpart in another possible world who loses. Under this solution to the problem of transworld identity, any individual in any possible world is worldbound. Nixon himself inhabits one and only one world: the actual world. Unactualized possible properties for Nixon are construed as the having of those properties by Nixon's counterparts in other possible worlds.[19] And here, once again, we might consider two different attitudes towards the Lewisian construal, a realist subsidiary attitude A_3 and an instrumentalist subsidiary attitude A_4.

According to A_3, for the actuality-bound Nixon to possibly have the property of losing the 1968 election *just is* for a counterpart of Nixon, Nixon*, to have the property in the counterfactual situation to which Nixon* is bound. This is what it turns out to be for an individual to possibly have a property lacked in actuality upon closer theoretical scrutiny. According to subsidiary instrumentalist attitude A_4, by contrast, the account of Nixon's possible loss in terms of Nixon* losing represents how the actuality-bound Nixon can have the property in another possible world ("in absentia", so to speak[20]). Nixon*, while flesh and blood (we suppose), is a representation of Nixon within the overall explanation of the fact that Nixon satisfies the property of losing the 1968 election in absentia. Nixon* losing is not itself expected to reveal the nature of Nixon's possible loss.

So much for general attitude *A* towards the possible world construal of *de re* modality and the concomitant problem of transworld identity. This general attitude is a realist attitude that considers a possible world portrayal of Nixon's possible loss as what the relevant modal fact turns out to be upon closer theoretical scrutiny. But we saw that even under the auspices of realist attitude *A* there might still be room for instrumentalist subsidiary attitudes A_2 and A_4, local instrumentalisms under a generally realist attitudinal umbrella.

An alternative general attitude towards the possible world construal of *de re* modality, attitude *B*, is thoroughly instrumentalist. The possible world construal is meant to represent the fact that Nixon might have lost despite actually winning. But it is not as though what it is for the *de re* modal fact to obtain *just is* for there to be a possible world according to which Nixon loses. The possible world construal plays a certain role within an overall explanation of how actual things might have had properties they do not in fact possess.[21] The explanatory utility of possible worlds for modal metaphysics easily outstrips the value of identifying modal facts with their possible world surrogates. The possible worlds apparatus helps represent in a more holistic manner how possibilities are connected with one another – how, say, Nixon's possible loss is connected with Humphrey's possible win – and how possibilities are connected with facts about actuality with certain modal implications – how, say, Nixon's possible loss is connected with the fact that the 1968 election was fair. Such explanatory tasks can be met without treating the representation of Nixon's possible electoral loss in terms of possible worlds as revealing the underlying nature of the fact in question, as a realist about the possible world portrayal would have it. *Pace* realism, possible worlds can play a useful modeling function without telling us with regard to each modal fact what makes it the fact that it is. What makes each such fact the fact that it is need not have anything to do with possible worlds.[22]

[19] We ignore the further Lewisian emphasis on the inconstancy of counterparthood.

[20] See Lewis (1986: 9–10).

[21] It might be maintained, for example, that things possibly have just those properties tolerated by what the things are in the most demanding sense – by their natures or essences. See my (2012: Chs. 1–2) for one development of this line of thought. A possible world construal could still act as a heuristic for such an account.

[22] Attitude *B* is exhibited throughout Kripke (1980) and leads Kripke to regard the problem of transworld identity as a pseudo-problem engendered by the wrong attitude towards possible worlds.

4 Frege on Indirect Reference

A third and final illustration of the choice between realist and instrumentalist attitude towards theoretical representations within philosophical explanation is provided by a central tenet of Frege's philosophy of language: the theory of indirect reference. Here a default realist attitude has had far reaching implications for the metaphysics of mind.

Very briefly, Frege's theory of sense and reference offers a two-tiered account of semantic significance in response to a perceived need for semantic analysis to explain our epistemic rapport with linguistic expressions alongside other familiar explananda, such as productivity. If we take the contributions of sub-sentential expressions to the semantic significance of whole sentences to be exhausted by whatever the sub-sentential expressions refer to, then under commonplace assumptions a true sentence of the form $a = b$ will seem to have the very same semantic significance as a correlative sentence of the form $a = a$. And yet our epistemic rapport with sentences of the form $a = b$ can differ greatly from our epistemic rapport with correlative sentences of the form $a = a$.[23]

Frege's well-known response to this observation – bred by the conviction that it falls within the purview of a semantic analysis to explain our epistemic rapport with linguistic expressions – is to associate semantically significant units not only with a reference but also with a sense, a "mode of presentation" of the reference. So while a and b are co-referential as demanded by the truth of a sentence of the form $a = b$, they may be associated with distinct senses. And so, the overall semantic significance of a true sentence of the form $a = b$ may be different from that of a correlative sentence of the form $a = a$. Frege's term for the senses expressed by whole sentences, which are composed of the senses of the sub-sentential expressions, is *Gedanken* – thoughts. They are the objects of our intellectual grasp and the modes of presentation of the things to which our sentences refer, the True and the False.

With the apparatus of sense and reference in hand, we may now consider the theory of indirect reference. The original framework explains how the sentences 'Danzig is pretty' and 'Gdansk is pretty' might differ in overall semantic significance despite the co-referentiality of the names 'Danzig' and 'Gdansk'. This might explain how Hilary can hold the sentence 'Danzig is pretty' to be true while failing to hold the sentence 'Gdansk is pretty' to be true.[24] The explanation is that 'Danzig' and 'Gdansk' express different senses; so the whole thoughts expressed by 'Danzig is pretty' and 'Gdansk is pretty' are different; and so, Hilary can believe that Danzig is pretty without believing that Gdansk is pretty. It is this last move that merits further scrutiny.

According to Frege's theory of indirect reference, while the sentences 'Danzig is pretty' and 'Gdansk is pretty' are true or false together, the sentences 'Hilary believes that Danzig is pretty' and 'Hilary believes that Gdansk is pretty' can easily diverge in truth-value. What determines that the first belief report is true while the second false are facts surrounding Hilary's beliefs. Here is the situation according to the proposed semantics. Begin with the simple sentences 'Danzig is pretty' and 'Gdansk is pretty'. 'Danzig' refers to a certain city and 'is pretty' refers to a first-level concept $A(x)$ that maps pretty objects to the True and maps other objects to the False. The sentence 'Danzig is pretty' is true to the extent that the city as argument for the concept $A(x)$ yields the True as value. Assuming that 'Gdansk' refers to the same thing as 'Danzig', the second sentence is true to the same extent as the first. But the sentences differ in the thoughts they express to the extent that 'Danzig' and 'Gdansk' differ in the senses they express, modes of presentation of one and the same city. Now consider the belief report 'Hilary believes that Danzig is pretty'. The name 'Hilary' refers to Hilary and expresses a suitable sense, a mode of presentation of the man. But the name 'Danzig' within the clausal complement of the belief report refers not to its ordinary reference, the city, but rather to its "indirect" reference, which is the sense of 'Danzig' in the simple sentence 'Danzig is pretty', a mode of presentation of the city.[25] Similarly, the first-level predicate 'is pretty' within the clausal complement of the belief report refers not to its ordinary reference, $A(x)$, but rather to its indirect reference, which is its ordinary sense, i.e. the sense of 'is pretty' in the simple sentence 'Danzig is pretty', a mode of presentation of the first-level concept. And the sentence 'Danzig is pretty' as the clausal complement of the belief report refers not to its ordinary reference, the truth-value, but to its indirect reference, which is its ordinary sense, i.e. the thought that Danzig is pretty, a mode of presentation of the truth-value. Accordingly, the dyadic predicate 'believes' refers to a dyadic relational concept $B(x, y)$ that maps believers and thoughts believed to the True, otherwise mapping pairs of objects to the False. It is thus that the sentence 'Hilary believes that Danzig

[23] The classic statement of the problem is the opening paragraph of Frege (1948).

[24] In discussion Hilary Putnam once expressed astonishment at the fact that Danzig is Gdansk.

[25] The indirect *sense* of 'Danzig' is the mode of presentation of the indirect reference and is other than the sense of 'Danzig' in the simple sentence, but we leave indirect senses aside. For further discussion of this issue see my (2018).

is pretty' can be true while the sentence 'Hilary believes that Gdansk is pretty' is false. While 'Danzig' and 'Gdansk' are co-referential in the true identity 'Danzig is Gdansk', 'Danzig' in the first belief report need not be co-referential with 'Gdansk' in the second belief report. The first report truly relates Hilary to the thought that Danzig is pretty. The second falsely relates Hilary to the thought that Gdansk is pretty.

Let us now step back from these details and consider the situation afresh. We wanted to explain how it is that Hilary believes that Danzig is pretty while failing to believe that Gdansk is pretty despite Danzig and Gdansk being one and the same city. Semantically ascending, we turned our attention to how it is that 'Hilary believes that Danzig is pretty' is true while 'Hilary believes that Gdansk is pretty' is false despite the truth of 'Danzig is Gdansk'. Frege provides a semantic apparatus, the theory of indirect reference, that delivers said result by taking belief reports to relay a relation of belief that believers bear to thoughts believed. The theory offers an elegant semantic treatment of belief reports.[26] Semantically descending and going back to the specific case before us, Frege's semantic analysis represents how Hilary can believe that Danzig is pretty while failing to believe that Gdansk is pretty. The analysis in terms of the belief relation relating the believer to the thought believed is a representation wielded in the explanation of the facts surrounding Hilary's cognitive situation. But it is all too common to regard this representation realistically as revealing what it is for Hilary to believe what he does. Indeed, the metaphysics of mind has often considered it a *datum* that cognitive states such as belief – so-called propositional attitudes – are relations cognitive agents bear to the contents of whole declarative sentences (often construed via more direct relations to and from things whose contents are the contents of whole declarative sentences).[27] An instrumentalist attitude, on the other hand, regards this as no datum. Relations of agents to the contents of whole declarative sentences are representations adduced for particular explanatory purposes and are not to be regarded as revealing the nature of the cognitive facts themselves.[28] The kind of explanation for which these relations were originally adduced was a semantic account of how the significance of reports of belief and other cognitive attitudes depends on the significance of their parts and their mode of composition. Treating the pronouncements of Frege's theory of indirect reference realistically within the metaphysics of mind depends on (a) treating them realistically within a *semantic* explanation of what it is for belief reports to mean what they do, while (b) treating the significance of the reports as revelatory of the nature of the facts of belief. It is thus that a semantic analysis of a belief report is taken to reveal the nature of the cognitive fact being reported. An instrumentalist attitude resists this two-step procedure.

We note that an instrumentalist attitude in this case need not involve any firm conviction that belief and other cognitive states will ultimately *not* turn out to be relational in the way envisaged by literalizers of Frege's semantic apparatus of indirect reference. Down the line there might be substantive empirical reasons in favour of some such relational story. But these are early days of cognitive theorizing. From an instrumentalist standpoint, ready inference from the details of Frege's semantic apparatus to the reality of the significance of belief reports, and then to the reality of beliefs themselves, is done at our peril. The theory of indirect reference captures what we say when we report beliefs by way of generating truth-conditions for our reports. An instrumentalist attitude would resist reading off from the theory a metaphysics of attitudes.

5 Semantics as Metaphysics?

Working backward from the example of the theory of indirect reference to the earlier one, the possible world construal of *de re* modality exhibits a strikingly similar pattern. A semantic construal of the syntactic construction $\exists x \Diamond \phi x$ in the language of quantified modal logic – itself a representation of a fragment of natural language – is regarded realistically in the semantic treatment of the *de re* modal locution and then taken to reveal the nature of the *de re* modal fact. (In the Lewisian version the nature of the modal fact is revealed by the interpretation of the counterpart-theoretic translation of the quantified modal logic construction.) Our modal discourse discloses the nature of the underlying modal facts through the prism of the formal semantic apparatus of possible worlds.

Working backward still, the theme recurs in Frege's treatment of the numbers as logical objects with which we began. Frege's original construction is the outcome of a semantics-first procedure. After carefully examining everyday uses of numerical expressions Frege (1953) famously concludes that "the content of a statement of number is an assertion about a concept" (§46). He then proceeds to contemplate the idea

[26] But we set aside the issue of semantic innocence and its significance for semantic theorizing. Further features of this theory *qua* semantic theory are discussed in my (2018).

[27] Examples of such an approach abound, including Davies (1991), Lycan (1993), and Rey (1995).

[28] In Simchen (forthcoming) such an attitude is taken up and defended in some detail.

of numerical attributions as structureless second-level predicates – the adjectival strategy – which is subsequently claimed to be vulnerable to insurmountable difficulties.[29] Frege then turns his attention to the semantic treatment of sentences of the form 'The number of ϕs is the number of ψs'. The numeral 'four' is claimed to be a singular term even in such adjectival constructions as 'The King's carriage is drawn by four horses', which means that it refers to an object. The object in question is finally construed as an extension of a certain second-level concept, *equinumerous with* $\underline{x = \mathbf{0} \vee x = \mathbf{1} \vee x = \mathbf{2} \vee x = \mathbf{3}}$, which is itself the reference of the second-level predicate 'equinumerous with *F*' for any first-level predicate *F* that refers to a first-level concept with exactly four objects in its extension. The proposed semantic treatment of numerical discourse is taken to reveal the numerical facts themselves – that for the King's carriage to be drawn by four horses *just is* for the first-level concept *horse drawing the King's carriage* to be a member of the extension of the second-level concept *equinumerous with* $\underline{x = \mathbf{0} \vee x = \mathbf{1} \vee x = \mathbf{2} \vee x = \mathbf{3}}$. The nature of the numerical fact is thus revealed by the semantics of its report.[30]

Frege's attitude towards his construction of number is clearly realist in our sense, as are the possible world metaphysician's attitude towards the possible world capture of *de re* modality and the Fregean attitude towards the account of cognitive attitudes as relations to Fregean thoughts. In each case a realist might say that the theory offers a theoretical identification, much like the identification of water as hydrogen hydroxide or gold as the element with atomic number 79. It is gratuitous to suppose that being hydrogen hydroxide only represents water for some broader theoretical purpose or other. Hydrogen hydroxide is what water turns out to be upon closer theoretical scrutiny. Similarly, it is now claimed, *horse drawing the King's carriage* being a member of the extension of *equinumerous with* $\underline{x = \mathbf{0} \vee x = \mathbf{1} \vee x = \mathbf{2} \vee x = \mathbf{3}}$ is what the King's carriage being drawn by four horses turns out to be; Nixon losing the election in another possible world (or the counterpart Nixon* losing that election) is what possible loss for Nixon turns out to be; and Hilary bearing the attitudinal relation to the thought expressed by 'Danzig is pretty' while not bearing it to the thought expressed by 'Gdansk is pretty' is what Hilary believing Danzig to be pretty while not believing Gdansk to be pretty turns out to be. The instrumentalist, on the other hand, will point to the distinctly semantic origin of the accounts on offer and the distinctive theoretical aims of semantic explanation as compared with philosophical explanations of the facts reported by the factual reports. And so the question naturally arises: are there general guidelines for which attitude, realist or instrumentalist, is more apt for a proposed theoretical capture of a target subject matter in a given case?[31]

Consider familiar cases of theoretical identification such as water being hydrogen hydroxide or gold being the chemical element with atomic number 79. Theoretical identifications in natural science provide paradigm examples for when a realist attitude towards the representation is warranted. Hydrogen hydroxide really is water. Being hydrogen hydroxide is not only a representation of water at the service of some broader theoretical purpose – it is also what water turns out to be upon closer theoretical scrutiny. Indeed, a first salient feature of representations for which a realist attitude is clearly warranted is the presence of a pretension to uncover the underlying nature of the represented, what we might call *realist purport*. In representing gold as the element with atomic number 79 we aim to uncover the nature of gold. There are surely cases of representing gold for other theoretical purposes that lack such pretension. Consider again the representation of gold as a standard for pre-20th century monetary systems in the explanation of the economic workings of certain monetary systems. Gold is thus represented under a widely assumed instrumentalist attitude – to be the gold standard is not presumed to be part of what it is to be gold. The latter case is distinctly unlike the identification of gold as the element with atomic number 79 when it comes to realist purport.

A second important feature of representations for which a realist attitude is clearly warranted is that the surrounding theory does not require us in general to switch from the subject matters of our basic everyday claims to something other than what those claims are pre-theoretically about unless such switching is

[29] Of particular note here is the so-called Julius Caesar problem. See Heck (2011) for an extensive treatment of the issue.

[30] See Steiner (1995) for an interesting discussion of Frege's methodology in the context of problems of applicability in the philosophy of mathematics. Given my characterization of Frege's methodology as semantics-first, I disagree with Steiner's suggestion that Frege solves the metaphysical problem of how applied mathematics is possible. Under Frege's construal, when all is said and done mathematical applicability boils down to a matter of membership of a referent of a first-level predicate in the extension of a referent of a second-level predicate. This mislocates the problem of how abstracta can apply to worldly concreta as a problem to be solved within the background semantic apparatus. I defer further discussion of the issue for another occasion.

[31] By 'subject matter' here and elsewhere I mean the fact or facts portrayed by a claim or claims without further commitment to a specific theoretical articulation of the pre-theoretical notion (such as the one offered in Lewis (1988) or the more recent one offered in Yablo (2014)).

demanded by clear theoretical progress. Conservatism as to subject matter is a natural default.[32] Conservatism as to subject matter does not mean that properties previously unheeded and unexplained have not emerged in the course of identifying water as hydrogen hydroxide, say. But the subject matters of our basic pre-theoretical water-claims are recognizably what we initially set out to explain even in the advent of the theory. They are water-facts. There are surely cases that require us to revisit and revise subject matters of basic everyday claims. Pre-theoretically we distinguish thunder from lightning when we affirm such claims as that the lightning preceded the thunder or when we affirm that the thunder was loud and the lightning bright and deny that the lighting was loud and the thunder bright. It might thus seem that the thunder and the lightning are not at bottom one and the same event of electrical discharge. And yet that is exactly what it is. The right thing to say about such cases is that the identification of thunder and lighting as one and the same event of electrical discharge entails revision of pre-theoretical subject matter of everyday thunder-and-lighting discourse but that such revision is clearly warranted by gained theoretical dividends overall.

A third salient feature of representations for which a realist attitude is clearly warranted is that the facts of representation fall within the purview of the surrounding theory. How water is represented as hydrogen hydroxide or gold as the element with atomic number 79 are matters that are presumed to fall within our grasp due to our handle on how microstructure is linked to macro features in light of the achievement of physical chemistry. But this point extends beyond theoretical identification and beyond micro-structural representation. Consider the representation of nasonite ($Pb_6Ca_4Si_6O_{21}Cl_2$) as hexagonal-dipyramidal within crystallography. Being hexagonal-dipyramidal is a denomination originating from solid geometry. But how it is that nasonite is such is well understood given the chemical nature of the substance. The fact of the representation of nasonite as hexagonal-dipyramidal does not raise perplexities beyond those tackled by the surrounding theory. Crystallography details how it is that nasonite, given its chemical nature, should be thus represented geometrically.

We thus distill three conjectured necessary conditions for when a realist attitude towards a theoretical representation is warranted. The first is the presence of realist purport – a pretension to reveal the nature of the represented that is clearly lacking in the gold standard case, for example. A second is conservatism as to subject matter: the preservation of pre-theoretical subject matter – as in the water case – not come-what-may but in such a way as to be sensitive to substantial theoretical benefits gained by revision – as in the thunder/lightning case. A third condition is that the fact of representation should itself be covered by the surrounding theory.

Armed with these conjectured necessary conditions, let us now turn to representations wielded within philosophical explanation. Such representations often meet the first condition. Realist purport associated with particular theoretical representations in philosophy is prevalent, undoubtedly fuelled by the common understanding of the philosophical enterprise as entrusted with the task of revealing the nature of whatever falls within its purview.[33] But how about the other two conditions? According to the second condition the representations of philosophical explanation should not require substantial revision as to the subject matter of everyday claims unless such revision is warranted by clear theoretical benefits gained by the revision. According to the third condition these representations should apply to the represented in a manner that is itself well-understood in light of the surrounding theory. It now appears that in the three case studies we have been discussing the second and third conditions for when a realist attitude is warranted are not met.

Regarding the third condition, in particular, we have witnessed that in each case the proposed representation was originally devised to serve distinctly semantic aims within an explanation of the significance of our reports of relevant facts – arithmetical, modal, or cognitive. Take again the possible world capture of a *de re* modal fact, a piece of applied mathematics, and compare it to the applied mathematics in the nasonite case just mentioned. Why the possible world representation should apply to the modal fact that Nixon might have lost the 1968 election surely does not enjoy the transparency of the applicability of the solid

[32] Quine (1957) memorably puts the point as follows:

> We imbibe an archaic natural philosophy with our mother's milk. In the fullness of time, what with catching up on current literature and making some supplementary observations of our own, we become clearer on things. But the process is one of growth and gradual change: we do not break with the past, nor do we attain to standards of evidence and reality different in kind from the vague standards of children and laymen. Science is not a substitute for common sense, but an extension of it. (2)

[33] This is not to deny that some philosophers – Lewis chief among them – have resisted drawing a default implication from the global revelatory pretension of a philosophical theory to the more specific revelatory pretensions associated with particular theoretical representations deployed therein. The point in the text is diagnostic and targets a widespread tendency.

geometrical denomination of hexagonal-dipyramidality to nasonite. This is not to deny that the applicability of mathematics – of which the applicability of hexagonal-dipyrmidality to nasonite is a special case – may seem generally perplexing. But why, beyond the general perplexity of mathematical applicability, the formal semantic analyses of various factual reports should apply to the explananda of philosophical explanation as pertaining to the facts being reported is far from clear. They seem like alien transplants from a remote theoretical endeavour.

Regarding the second condition of preservation of pre-theoretical subject matter unless revision is demanded by clear theoretical dividends – conservatism as to subject matter – we register some further observations. Consider again arithmetical discourse. Numerical attributions figure prominently among basic claims we make in everyday life. Frege insists that the subject matter of 'The King's carriage is drawn by four horses' is not the carriage and the plurality of horses drawing it being four, as we might ordinarily expect, but a first-level concept *horse drawing the King's carriage* belonging to the extension of the second-level concept *equinumerous with* $\underline{x} = \mathbf{0} \vee x = \mathbf{1} \vee x = \mathbf{2} \vee x = \mathbf{3}$. The same revisionism as to subject matter is exhibited vividly regarding such universal claims as 'All whales are mammals', which Frege claims to be about concepts rather than animals.[34] The Fregean analysis effectively swaps the everyday subject matter – the carriage being drawn by a plurality of horses numbering four – for a theoretical proxy – one concept belonging in the extension of another. Or consider again the *de re* modal claim that Nixon might have lost. Under one realist construal possible loss for Nixon is loss for Nixon*. The subject matter of the everyday claim that Nixon might have lost, a basic everyday modal claim concerning Nixon, is swapped for another, a fact concerning a counterpart.[35] Or consider, finally, the everyday attribution that Hilary believes Danzig is pretty while not believing Gdansk is pretty. The pre-theoretical subject matter is Hilary's beliefs. And yet relations to abstracta – Fregean thoughts – are again something else altogether.[36]

Now, theoretical advances can surely exact the toll of revision when it comes to subject matters of basic pre-theoretical everyday claims, as noted above. A familiar example is Putnam's (1975) jade case, which upon closer theoretical scrutiny was revealed to be not one but two distinct minerals, jadeite and nephrite. That the jade-phenomena lack the unity we were pre-theoretically inclined to confer upon them is a price incurred by theoretical progress.[37] So in keeping with the second conjectured necessary condition for when a realist attitude is warranted, we need to ask about the cases we have been considering whether revisions of pre-theoretical subjects matters of basic everyday claims are warranted by significant theoretical achievement. A radical departure from pre-theoretical subject matters of basic everyday claims that is not accompanied by clear theoretical progress should rouse the suspicion that a representation has taken on a life of its own. It should certainly incline us to resist the tendency to regard the representation realistically. In keeping with a general conservatism as to subject matter, I conclude that such is the case in all three areas we have been discussing – arithmetical facts, *de re* modal facts, and cognitive attitudinal facts. The revision of pre-theoretical subject matters of basic everyday claims, coupled with scant evidence for clear theoretical dividends demanding such revisionism, should disincline us to regard the relevant theoretical representations realistically. The representations of these facts within the relevant philosophical theories should not be regarded as revealing the nature of the represented facts.

6 Metaphilosophical Instrumentalism

We began by considering in general terms the choice of attitude towards representations deployed within philosophical explanation. We then looked at some familiar cases outside philosophy for which a realist attitude seems clearly warranted and identified conjectured necessary conditions for such warrant. We concluded that for the philosophical theoretical representations discussed above the second and third conditions aren't met. Regarding the third condition, in particular, we noted that in all the cases discussed it appears that a realist attitude treats semantic representations originally adduced to explain the significance of certain factual reports as revealing the nature of the facts being reported. Let us now draw some broader implications for general philosophical methodology.

[34] See Frege (1953: §47).

[35] This seems to be the crux of Kripke's so-called Humphrey objection against Lewis's Counterpart Theory. See Kripke (1980: 45 n.13).

[36] In Simchen (2012: Ch. 5) I explore this particular distinctness in some detail.

[37] Unlike the lightning/thunder case, where apparently bifurcated phenomena are revealed by our theories to be unified, in the jade case apparently unified phenomena are revealed by our theories to be bifurcated. For an illuminating discussion of the jade case, see Hacking (2007).

Any teacher of philosophy is familiar with the incredulity exhibited by students regarding theoretical constructions representing pre-theoretical subject matters within putative philosophical explanations. "Numbers are sets." – "But we don't count with sets!" (the reaction voicing perhaps the sentiment that numbers are inherently quantitative while sets are not). "Possibilities are possible worlds." – "But what do possible worlds have to do with *the world*?" (voicing perhaps the worry that unactualized possibilities are aspects of actuality whereas non-actual possible worlds are alternatives to actuality). "Beliefs are relations to the semantic contents of whole declarative sentences." – "But believing doesn't need to involve language!" (voicing perhaps the thought that mental states such as belief are exhibited by non-linguistic animals too). "God is that than which nothing greater can be thought." – "But what does our ability to think have to do with what God is?" (voicing perhaps the concern that the nature of the divine should float free from matters of conceivability). In all these exchanges the student may be expressing a correct idea under a certain reading of the teacher's claim. The teacher, on the other hand, need not be committed to the ascribed reading. The student complains that being such that nothing greater can be thought isn't what God really *is* at bottom. The teacher could reply:

> Being such that nothing greater can be thought is a representation of God within an ingenious argument for God's existence. Anselm may have thought that being such that nothing greater can be thought reveals the nature of the divine. We need not assume that it does in order to appreciate the argument's achievement. The representation of God as that than which nothing greater can be thought is still serviceable within the broader effort to show how God's existence is the sort of thing that might be proved.

The teacher's point can also be put like this: Even if we assume that a particular theoretical representation (e.g. Anselm's definition of God) deployed within an overall explanatory effort (the ontological argument) doesn't reveal the nature of what it represents (God), it can still have a more general or holistic explanatory utility (illustrating how something like God's existence can admit of proof in the first place).[38]

Philosophical explanation is a human enterprise, not an exercise in divine revelation. It deploys representations of its various subject matters in an effort to shed light on them. But reasons for deploying theoretical representations within our various explanations are many and varied. We philosophers might imagine our representations as tapping into the nature of what they represent. We might imagine ourselves to be undertaking the physical chemistry of reality, so to speak, taking each of our representations as offering something on par with the representation of gold as the element with atomic number 79 or of water as hydrogen hydroxide. In point of fact, even if there is something to this aspiration on a larger scale, for a great many cases of philosophical explanation there is very little to support the contention regarding the particular theoretical representations deployed therein, or so I have argued. Theoretical representations deployed in philosophy are often closer to such cases as representing the meaning of a sentence as its truth-condition in truth-conditional semantics. In the latter case it is not supposed by the working semanticist that the truth-condition is what the meaning of the sentence turns out to *be* upon closer theoretical scrutiny, along the lines of water turning out to be hydrogen hydroxide. Rather, the meaning of the sentence is so represented within a more holistic effort to explain the compositionality of meaning – how the meanings of complex expressions depend on meanings of their parts – which in turn might ultimately contribute to the theoretical exploration of our capacity to produce and understand novel sentences. That the meaning of a sentence should be theoretically identified with its truth-condition is no part of such broad explanatory aims.

Attending to the distinction between realist and instrumentalist attitudes towards theoretical representations allows us to consider the question of aptness or warrant when it comes to how we should regard theoretical representations within various philosophical explanations. These issues fall under general philosophical methodology, under what is also known as *metaphilosophy*. There seems to be a prevalent attitude in philosophy whereby the representations deployed in philosophical explanation are to be regarded under a realist attitude by default. Call this default inclination to adopt a realist attitude towards the theoretical representations of philosophical explanation *metaphilosophical realism*. Metaphilosophical realism is a higher-order attitude – an attitude towards attitudes towards theoretical representations in philosophy.

[38] I am ignoring here well known problems with Anselm's argument, the most striking perhaps is the "exportation" fallacy pointed out by Gaunilo (albeit in embryonic form) – that from the fool's concession that it is thought that: the *x* such that for all *y* it is impossible to think that $y > x$ exists, it doesn't follow that the *x* such that for all *y* it is impossible to think that $y > x$ is also such that it is thought that: *x* exists.

Under metaphilosophical realism a philosophical account of some subject matter *X*, ushering in a theoretical representation of *X*, will invariably tell us what *X* is in the most demanding sense. The situation here is assumed to be not unlike the physical-chemical representation of gold as atomic element Au, where being Au tells us what gold really is at bottom, what we've all been speaking of all along in speaking of gold.

Against metaphilosophical realism we have the higher-order attitude we can now call *metaphilosophical instrumentalism*. The metaphilosophical instrumentalist will not treat a theoretical representation of *X* within some putative philosophical explanation, however successful, as revealing the nature of *X* by default. Metaphilosophical instrumentalism proclaims that an instrumentalist attitude towards theoretical representations in philosophy can be warranted. We aren't compelled to regard philosophy's representations realistically. This attitude comes in different grades of strength. An extreme version denies that a realist attitude towards theoretical representations in philosophy is *ever* warranted. A more moderate version will leave room for the possibility of a warranted realist attitude towards some such representations. I have had little to say here to decide among such alternatives. My main efforts have been directed at arguing against metaphilosophical realism. The extreme version of metaphilosophical instrumentalism does seem doubtful, however. In proclaiming that a realist attitude towards theoretical representations deployed in philosophy is never justified, the instrumentalist is either giving voice to a prior conviction that a realist attitude towards theoretical representations *tout court* is never justified – a conviction I do not share; or else the instrumentalist is presupposing a criterion for distinguishing philosophical explanation from other forms of explanation – a criterion whose existence I very much doubt.

Metaphilosophical instrumentalism incurs a special explanatory burden. If theoretical representations do not themselves reveal the nature of whatever they represent, how can an entire theoretical system to which they belong succeed in its explanatory aims? The answer is that such a system can reveal something important about the nature of its overall subject matter without each of its "moving parts" revealing the nature of what it represents individually. Humdrum cases of such disparity between representational wholes and their parts abound. Think of a plastic model of a molecule, with rods connecting plastic spheres representing individual atoms. While the model as a whole can reveal an important structural aspect of whatever it represents, the individual plastic spheres and rods do not. For a less cartoonish example, see the discussion of neologicism at the end of Section 2. Theoretical representations may reveal the nature of the represented collectively without doing so severally. Nature disclosure can occur at a macro-level for a system of theoretical representations without being achieved at the micro-level, one representation at a time.

Acknowledgements

For helpful discussion of this material I am indebted to audiences at Hebrew University, UC Santa Cruz, Tel Aviv University, and the University of Porto. For further comments I am indebted to Roberta Ballarin, Avner Baz, Jordan Dopkins, Naama Friedmann, David Kashtan, Aviv Keren, Samantha Matherne, Adam Morton, Nico Orlandi, Carl Posy, Diana Raffman, Paul Roth, Gil Sagi, Abe Stone, Charles Travis, John Woods, and several referees. I gratefully acknowledge the support of the Social Sciences and Humanities Research Council of Canada (Grant 435-2017-0133).

Competing Interests

The author has no competing interests to declare.

References

Davies, M. 1991. Concepts, Connectionism, and the Language of Thought. In: Ramsey, W, Stich, SP and Rumelhart, DE (eds.), *Philosophy and Connectionist Theory*. Hillsdale, NJ: Lawrence Erlbaum.

Dummett, M. 1993. *Origins of Analytical Philosophy*. Cambridge, MA: Harvard University Press.

Frege, G. 1948. Sense and Reference. *Philosophical Review*, 57: 209–230. DOI: https://doi.org/10.2307/2181485

Frege, G. 1952. On Concept and Object. *Mind*, 60: 168–180. DOI: https://doi.org/10.1093/mind/LX.238.168

Frege, G. 1953. *The Foundations of Arithmetic*. Oxford: Basil Blackwell.

Frege, G. 2013. *Basic Laws of Arithmetic*. Oxford: Oxford University Press.

Geach, P. 1955. Class and Concept. *Philosophical Review*, 64: 561–570. DOI: https://doi.org/10.2307/2182633

Hacking, I. 2007. The Contingencies of Ambiguity. *Analysis*, 67: 269–277. DOI: https://doi.org/10.1093/analys/67.4.269

Heck, R. 2011. *Frege's Theorem*. Oxford: Oxford University Press.

Kaplan, D. 1975. How to Russell a Frege-Church. *Journal of Philosophy*, 72: 716–729. DOI: https://doi.org/10.2307/2024635

Kripke, S. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Lewis, D. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Lewis, D. 1988. Statements Partly About Observation. *Philosophical Papers*, 17: 1–31. DOI: https://doi.org/10.1080/05568648809506282

Lycan, W. 1993. A Deductive Argument for the Representational Theory of Thinking. *Mind and Language*, 8: 404–422. DOI: https://doi.org/10.1111/j.1468-0017.1993.tb00292.x

Parsons, C. 1965. Frege's Theory of Number. In: Black, M (ed.), *Philosophy in America*. Ithaca, NY: Cornell University Press.

Plantinga, A. 1976. Actualism and Possible Worlds. *Theoria*, 42: 139–160. DOI: https://doi.org/10.1111/j.1755-2567.1976.tb00681.x

Putnam, H. 1975. The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science*, 7: 215–271. DOI: https://doi.org/10.1017/CBO9780511625251.014

Putnam, H. 1977. Models and Reality. *Journal of Symbolic Logic*, 45: 464–482. DOI: https://doi.org/10.2307/2273415

Putnam, H. 1978. *Meaning and the Moral Sciences*. Oxford: Routledge.

Putnam, H. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511625398

Quine, WV. 1951. Two Dogmas of Empiricism. *Philosophical Review*, 60: 20–43. DOI: https://doi.org/10.2307/2266637

Quine, WV. 1957. The Scope and Language of Science. *British Journal for the Philosophy of Science*, 8: 1–17. DOI: https://doi.org/10.1093/bjps/VIII.29.1

Reck, E. 2007. Frege-Russell Numbers: Analysis or Explication? In: Beaney, M (ed.), *The Analytic Turn: Analysis in Early Analytic Philosophy and Phenomenology*. New York: Routledge.

Rey, G. 1995. A Not "Merely Empirical" Argument for a Language of Thought. *Philosophical Perspectives*, 9: 201–222. DOI: https://doi.org/10.2307/2214218

Simchen, O. 2012. *Necessary Intentionality: A Study in the Metaphysics of Aboutness*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199608515.001.0001

Simchen, O. 2013. The Barcan Formula in Metaphysics. *Theoria*, 78: 375–392. DOI: https://doi.org/10.1387/theoria.6918

Simchen, O. 2017. *Semantics, Metasemantics, Aboutness*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780198792147.001.0001

Simchen, O. 2018. The Hierarchy of Fregean Senses. *Thought*, 7: 255–261. DOI: https://doi.org/10.1002/tht3.394

Simchen, O. (forthcoming). Instrumentalism About Structured Propositions. In: Tillman, C (ed.), *The Routledge Handbook of Propositions*. New York: Routledge.

Steiner, M. 1995. Applicabilities of Mathematics. *Philosophia Mathematica*, 3: 129–156. DOI: https://doi.org/10.1093/philmat/3.2.129

Wright, C. 1983. *Frege's Conception of Numbers as Objects*. Aberdeen: Aberdeen University Press.

Yablo, S. 2014. *Aboutness*. Princeton: Princeton University Press. DOI: https://doi.org/10.23943/princeton/9780691144955.001.0001

How to cite this article: Simchen, O. 2019. Realism and Instrumentalism in Philosophical Explanation. *Metaphysics*, 2(1), pp. 1–15. DOI: https://doi.org/10.5334/met.20

Submitted: 20 March 2019 **Accepted:** 02 August 2019 **Published:** 11 September 2019