Towards Ideal Understanding

Mario Hubert
The American University in Cairo
Federica Malfatti
University of Innsbruck

October 12, 2022

Forthcoming in *Ergo*

What does it take to understand a phenomenon ideally, or to the highest conceivable extent? In this paper, we answer this question by arguing for five necessary conditions for ideal understanding: (i) representational accuracy, (ii) intelligibility, (iii) truth, (iv) reasonable endorsement, and (v) fitting. Even if one disagrees that there is some form of ideal understanding, these five conditions can be regarded as sufficient conditions for a particularly deep level of understanding. We then argue that grasping, novel predictions, and transparency are not reasonable conditions for ideal understanding.

Table of Contents

1 Introduction

Science aims at providing us with an understanding of reality. Scientists are currently trying to understand, e.g., the spread of COVID-19, the rise in temperature on our planet, the structure of dark matter, the phenomenon of antibiotic resistance. But how exactly does this work? How and when does science succeed in providing us with an understanding of reality? When does a phenomenon count as (scientifically) understood? What is (scientific) understanding?

Understanding is not an all-or-nothing matter. It admits of degrees. We can understand phenomena more or less, better or worse, superficially or in-depth, to some extent or fully. This seems to be the case even though we sometimes meaningfully use the verb "to understand" categorically, as in "You simply don't understand" or "Now I (finally) understand!". The claim that understanding admits of degrees is compatible with the idea that there is some threshold of minimal understanding that must be reached for attributions of understanding (such as "S understands P") to come out true. Once we have reached the threshold of minimal understanding, we can always get above it. That is, we can always improve, fine-tune, advance, or deepen our understanding, until we reach full (or at least a very high degree of) understanding.

Many authors try to shed light on understanding by specifying minimal conditions that must be fulfilled for attributions of understanding to come out true. Drawing on Kelp (2015), we follow a different strategy. We try to shed light on understanding by asking what it takes to understand phenomena *to the maximum conceivable extent*. In other words, we explore and analyze the state of *ideal understanding* and specify the conditions that must be fulfilled to reach such a state.

But why should one care about the state of ideal understanding at all? In practice, understanding is not an ideal. Understanding is not merely something that will be achieved at the very end of our scientific endeavors (if there ever will be such a thing), when all the questions about reality will have been answered. A significant degree of understanding has already been achieved in real-life scientific practice. Our current most successful scientific theories (and maybe also some of the scientific theories of the past that we left behind and replaced by better ones) succeed (or succeeded) in providing scientists with a *certain degree* of understanding of reality. Here is how Elgin (2017, pp. 15–16) puts it:

> I take it that science provides an understanding of the natural order. By this I do not mean merely that an *ideal* science would provide such an understanding or that at the end of inquiry science will provide one, *but that much actual science has done so and continues to do so*. […] So an adequate epistemology should explain what makes good science cognitively good. (Our emphasis)

However, we think that exploring the state of ideal understanding is a way to achieve precisely what Elgin urges epistemology to do: to explain what makes good science cognitively good. More precisely, by exploring the state of ideal understanding, we gain the conceptual resources necessary to explain how and to what extent our best science succeeds (or succeeded) in providing us with (a certain degree of) understanding of reality, even if ideal understanding might be a practically unreachable epistemic state.[1]

One may argue here that to achieve this goal we do not actually need to analyze the state of ideal understanding. To shed light on degrees of understanding, we should rather start by specifying what makes particular instances of understanding better or worse.[2] While we take this to be a promising perspective of inquiry, we think that our approach should be preferred because it has significant advantages. By focusing only on how understanding can be improved in a given context, one might overlook criteria for how understanding can be improved all things considered. This is why we decided to take a bird's eye view and to develop a framework

[1] Ideal understanding must not be confused with omniscience (see also Kelp, 2015). Omniscience is too idealistic an ideal. In his *Politics*, Aristotle refers to a similar hierarchy of ideals, in his case a hierarchy of ideal states. The highest ideal state would be a monarchy, where one ruler rules for the common good. Nevertheless, Aristotle warns us from pursuing such a state and so recommends a less ideal state, the polity, the rule of the majority for the common good. Omniscience would be something akin to Aristotle's view of monarchy, and ideal understanding something akin to polity. Ideal understanding is descriptive (that is, it describes to a certain extent what scientists do) and it is prescriptive (that is, it poses norms and guidelines for what scientist ought to do or can do under certain circumstances). Omniscience would be vacuously prescriptive because it would always demand that a scientist should know *more* but not exactly what this *more* amounts to.

[2] Thanks to an anonymous reviewer for raising this concern.

that incorporates all possible ways for how understanding can be enhanced. Hopefully, this will constitute the groundwork for exploring degrees of understanding (although we do not pursue this project in detail in this paper).

Moreover, and relatedly, if one started by trying to distinguish between better and worse understanding, one would run into the problem of specifying the relevant contextual factors for improving understanding. Whether one reaches the understanding threshold or not, and how much above the threshold one is located, depends on contextual factors. Who is the subject or potential understander? Is it a child, an expert, or a Nobel prize winner? Which goals and aims is the subject pursuing? This contextuality makes the project of directly developing a theory of understanding improvement rather complicated without having an overarching framework of ideal understanding— which is what we are advocating. If we ask what the perfect state of understanding would be like, we can abstract away from all these contextual factors. Of course, in the particular real-life instances of understanding these factors will re-enter and play a crucial role; consequently, there will probably be a variety of different ways in which the ideal state of understanding can be approximated. Exploring this variety of ways is certainly an important project, but it goes beyond the scope of this paper.

Furthermore, as we will discuss, prominent authors, such as Kelp and Khalifa, plausibly define a state of maximal understanding; our account of ideal understanding is intended to show how their accounts are exemplified in concrete scientific practice. After all, it is part of scientific practice to refer to ideals: What does an ideal quantum field theory look like? What would be an ideal study to determine successful medical interventions? An ideal gives scientists an orientation for action and inquiry. Even those who think of ideal understanding as metaphysically or epistemologically problematic can conceive of it, at the very least, as a helpful construct for exploring how concrete instances of understanding can be evaluated.

2 Current Theories of Ideal Understanding and Their Problems

Kelp (2015) and Khalifa (2017) propose theories of maximal understanding and define degrees of understanding with reference to such a state. In this section, we briefly reconstruct their views and argue that, despite important merits, they face certain problems that our account solves. This will prepare the ground for our own theory of ideal understanding, that does not contradict, and but rather complements, Kelp's and Khalifa's theories.

2.1 Kelp's Maximal Understanding

Central to Kelp's theory (and to ours too) is the concept of phenomenon. The phenomenon, for Kelp, counts as the target or object of understanding. It is hard to precisely define what phenomena are, as many different events and processes qualify as phenomena. The most comprehensive analysis of phenomena has been conducted by Bogen & Woodward (1988) and Woodward (2011). They define phenomena as *features of the world that in principle could recur under different contexts or conditions*. They distinguish phenomena from *data*, which are *public records produced by measurement and experiment that serve as evidence for the existence or features of phenomena.* Although Kelp does not refer to these works, his characterization of phenomena, although rather in terms of examples than a general definition, seems to be consistent with them.

Kelp's goal is then to explain what it means for an agent to understand a phenomenon to the maximal possible degree. In order to do so, he introduces two conditions. First, the agent needs *fully comprehensive knowledge* about the phenomenon, that is, the agent needs to know everything there is to know about the phenomenon. Consider the phenomenon of a cannonball falling to the ground. There are different ways to explain this phenomenon. One way to explain it is by means of Newton's law of gravitation. Given the initial conditions of the cannonball (position, velocity, and mass), the law tells us how the cannonball falls to the ground, namely, on a straight trajectory with constant acceleration. But we can explain how the cannonball falls to the ground also by means of conservation of energy. The potential energy of the body at the beginning is transferred into kinetic energy when falling.

But as Kelp rightly points out, one may still lack a certain degree of understanding, even if one has fully comprehensive knowledge. In the case of the falling cannonball, one may know those two explanations without realizing that they are related: it is possible to derive the law of energy conservation from Newton's law of gravitation (Goldstein et al., 2001 Ch. 1.1). Therefore, Kelp also demands *maximally well-connected knowledge* of the phenomenon; that is, the agent needs to know how the explanations of the phenomena are related to one another (Is one entailed by the other? Are they independent?). Combining his two conditions, Kelp's account of maximal understanding amounts to (Kelp, 2015, p. 3811):

> **Maximal Understanding**: If one has fully comprehensive and maximally well-connected knowledge of a phenomenon P, then one has maximal understanding of P.

We think that Kelp delivers an important and, in our opinion, a correct general framework for maximal understanding. Nonetheless, the framework is too general in this form and a step too far from scientific practice. How does a scientist accomplish fully comprehensive knowledge? What is the role of a scientific theory in providing such knowledge? What criteria

does a theory need to fulfill for doing so? How does a scientist accomplish maximally well-connected knowledge? Do these connections only need to hold among explanations or propositions regarding the *particular* phenomenon? We think that the well-connectedness of knowledge needs to go farther than perceived by Kelp: our knowledge of a phenomenon must be also well-connected to our background beliefs. We may well have comprehensive and well-connected knowledge of a phenomenon, but this knowledge may not adequately match our background beliefs, which can be beliefs about metaphysics, about other scientific theories, or the practice of science. In this case, we suffer from some cognitive dissonance. To mitigate such dissonance, our knowledge about a phenomenon must be consistent and match with, what we call, the noetic system of an agent (we discuss this in detail in section 4.3).

2.2 Khalifa's Ideal Understanding

Inspired by Kelp's account, Khalifa (2017) develops an account of maximal understanding, which he calls *ideal understanding,* which shares some features of Kelp's maximal understanding but also differs in crucial respects. Khalifa (2017, p. 4) characterizes ideal understanding in the following way:

> **Ideal Understanding**: S ideally understands why P if and only if it is impossible for anyone to understand why P better than S.

In contrast to Kelp, Khalifa bases his ideal understanding on a comparison of degrees of understanding *between different agents*. Agent S has ideal understanding of a phenomenon P if no other agent (not only in practice but also in principle) can have a better understanding of P than S. How do we quantify that one agent *understands better* than another one? Khalifa builds his entire theory of understanding on this point. First, he tells us what minimal understanding is, and from there he builds up to ideal understanding by developing a criterion for *understanding better*.

> **Minimal Understanding**: S has minimal understanding of why P if and only if, for some Q, S believes that Q explains why P, and Q explains why P is approximately true (Khalifa, 2017, p. 14).

An agent S has minimal understanding of a phenomenon, if she has one approximately correct explanation of this phenomenon. Khalifa (2017, p. 7) demands the following four requirements for an explanation:

Q (correctly) explains why P if and only if:

(1) Q is (approximately) true;

(2) Q makes a difference to P;

(3) Q satisfies your ontological requirements (so long as they are reasonable); and

(4) Q satisfies the appropriate local constraints.

Khalifa's entire account is based on explanations, which some find controversial (Wilkenfeld, 2013, Kelp, 2015, Dellsén 2020). But given that one may base understanding on explanations, conditions (1) and (2) seem to be uncontroversial. Condition (3), on the other hand, might seem to be too strong a requirement and might also require a stance with respect to realism and antirealism (Chakravartty, 2017). In order not to be committed to a particular stance, Khalifa adds the qualifier "as long as they are reasonable". Whether demanding a certain (unobservable) ontology as part of an explanation depends on whether one regards such ontology reasonable or not, and this depends on where one stands in the realism-antirealism debate. Certain antirealists would abandon any ontological requirements on a scientific explanation, which would make condition (3) vacuous, and Khalifa embraces this option. Pleasing the antirealist with such a broad application of condition (3) makes Khalifa's theory more attractive to a broader group of scientists and philosophers. Nevertheless, regarding the goal of describing what ideal understanding amounts to, one may re-consider such a loose requirement regarding an unobservable ontology. Even if one may be skeptical with respect to the existence of unobservable entities, one may still grant that those entities deepen one's understanding of the phenomenon; t hey may help to foster one's intuition; or at least they may help memorize or visualize what is going on in the world. For example, mechanistic explanations, which depend in many cases on an unobservable mechanism, do provide a deeper understanding of a phenomenon than explanations that do not rely on such mechanisms (see Hubert, 2021 for this argument).

It is a bit unclear why Khalifa adds condition (4). He says that the first three conditions are global constraints that are valid for all kinds of scientific explanations, while condition (4) emphasizes that local constraints are needed too. The local constraints Khalifa has in mind are the specific requirements that are imposed on a scientific explanation *within a certain scientific discipline or context*. Some contexts or disciplines rely on idealizations, others demand causal explanations, etc. Khalifa wants to make sure that specific scientific contexts determine when an explanation is good enough or not (even when it fulfills the first three conditions).

Now having an account of scientific explanations and minimal understanding, we can continue to follow Khalifa in how he develops an account of *better understanding*. Roughly speaking, the more explanations an agent has at hand the better is her understanding (Khalifa, 2017, p. 14):

> **Better Understanding I**: S_1 understands why P better than S_2 if and only if:
> *Ceteris paribus*, S_1 grasps P's explanatory nexus more completely than S_2.

We agree with Khalifa that the more explanations we know about a phenomenon the better we understand it. Like in Kelp's case, Khalifa builds into his account a connectedness requirement by describing the set of explanations forming an interrelated *nexus*. And also, like

Kelp, Khalifa does not say if and how the explanatory nexus is anchored in other beliefs we form about the world.

Khalifa (2017, p. 14) presents another criterion, which may be powerful enough to say more about how the explanatory nexus is connected to an agent's background belief system:

> **Better Understanding II:** S_1 understands why P better than S_2 if and only if: Ceteris paribus, S_1's grasp of p's explanatory nexus bears greater resemblance to scientific knowledge than S_2's.

While Khalifa's first criterion describes *better understanding* "from below", that is, by building it up from grasping more explanations, the second criterion describes better understanding "from above" by focusing on how close the grasping of the nexus is to scientific knowledge. This characterization hinges, of course, on what Khalifa means by scientific knowledge. Khalifa does not really define scientific knowledge, and its relation to ideal understanding is not clear either. Although Khalifa writes, "Very roughly, ideal understanding is maximally scientific knowledge of a complete explanatory nexus (p. 15), it is unclear to us what "maximally" scientific knowledge is. It seems that scientific knowledge comes in degrees. Or Khalifa identifies ideal understanding with scientific knowledge simpliciter, and maximally scientific knowledge refers to the grasping of the explanatory nexus.

Anyway, we think that the notion of scientific knowledge may be able to comprise the connectedness requirement that we miss in Khalifa's first criterion. He describes that an agent gains scientific knowledge when she evaluates different possible scientific explanations for a particular phenomenon P. In a first step, he considers different candidate explanations and evaluates which are plausible and which are implausible. Although not explicitly mentioned by Khalifa, an agent may use her background beliefs in deciding whether an explanation is plausible or not. Khalifa briefly gives the example that one would immediately rule out an explanation of Newton's death "by appeal to alien laser guns". Without conducting an autopsy (which is now impossible anyhow) or historical research about the circumstances of Newton's death, one can refer to one's background beliefs about how people die and the likelihood that aliens have visited the Earth to render such an explanation implausible. One crucial ingredient in our account of ideal understanding is exactly this connection with our background beliefs.

Another important criticism against Khalifa's theory of understanding raises the issue that he ignores too much of the practical skills of a scientist and that he over-emphasizes the importance of explanations (de Regt & Baumberger, 2020; de Regt & Höhl, 2020). De Regt's contextualist theory of understanding is supposed to amend this problem, but we will see that he ultimately over-emphasizes the intelligibility of a scientific theory.

3 Searching for Conditions for Ideal Understanding

We will now discuss the main ideas of de Regt's contextualist theory of understanding (de Regt, 2017) as a counter-theory to Khalifa's explanationist theory. We will then discuss Wilkenfeld's multi-dimensional (MUDy) account of understanding (Wilkenfeld, 2017), which will be the anchor for our own account.

3.1 De Regt's Contextualist Theory of Understanding

Central to de Regt's theory of understanding is a scientific theory by which a scientist gains understanding by grasping an explanation provided by this theory (de Regt 2017). To provide understanding, the scientific theory must instantiate three qualities; it must be:

i) intelligible (to the scientist),
ii) empirically adequate,
iii) internally consistent.

But what does it take for a theory to be "intelligible", in de Regt's sense? De Regt defines intelligibility in a pragmatic way: a theory is said to be intelligible if a scientist can easily apply the theory for practical matters (de Regt, 2017, pp. 40 and 101). Therefore, the implementation of intelligibility depends on the theory and on what the scientist intends to accomplish with this theory. An instantiation of de Regt's pragmatic intelligibility criterion that is particularly suited for physics says that a scientist should be able to intuitively apply the theory *without making detailed calculations*. Since this kind of intelligibility does not only depend on the theory itself but also on the abilities and background knowledge of the scientist (and maybe also on the current state of technology for applying the theory), intelligibility is a non-intrinsic, contextual property of a theory. And because intelligibility in addition depends on the scientist using the theory, a theory that is intelligible to one scientist may be unintelligible (or less intelligible) to another.[3]

We agree with de Regt's contextual theory of understanding that successfully applying a theory and having strong intuitions about its consequences, granted the empirical adequacy and consistency requirement, typically results in a certain degree of understanding of the

[3] De Regt & Gijsbers (2017) have made the contextual theory even more pragmatic and turned it into a theory of *effectiveness*. They no longer define understanding relative to a theory but relative to a *representational device*, which can be specified in terms of "theories, models, and diagrams" (p. 50). The representational device need *not* be internally consistent, but only intelligible (as previously defined) and reliably successful. A representational device can be reliably successful in three different ways: (i) by making correct predictions, (ii) by guiding practical applications, (iii) by developing better science. Finally, a representational device is said to be *effective* if it is intelligible and reliably successful.

physical world; however, we think that de Regt's notion of intelligibility incorporates both too much and too little. It incorporates an aspect of understanding that goes beyond the scientist's skill to use the theory, on the one hand; and on the other, it does not offer the whole story about intelligibility because it disregards crucial aspects of it (particularly, when we aim to work out an ideal).

For de Regt, the pragmatic virtues of a theory do not only depend on the theory but also on the scientist using the theory. Different scientists may have different background knowledge, metaphysical commitments, etc. (de Regt, 2009, p. 592). And because they differ in these commitments, they value different virtues of theories (or even entire theories) differently. De Regt (2017, Ch. 5) delves deeper into the metaphysical commitments of physicists (he discusses Newton's theory of gravity in its historical development, which we will take up in section 4.3), and he distinguishes two kinds of intelligibility: *scientific intelligibility* and *metaphysical intelligibility*. Scientific intelligibility is achieved through using the tools of the theory to find adequate explanations and predictions. Metaphysical intelligibility is achieved through a metaphysical worldview that makes sense to the scientist and that often provides conceptual tools for scientific intelligibility.[4] Putting metaphysical intelligibility and scientific intelligibility under the same umbrella, de Regt puts too much weight on the significance of metaphysics for the usability and application of scientific theories. For, we can take a theory to be scientifically intelligible that we know to be wrong and implausible, given what we already hold true about reality. For example, one may even become an expert in the domain of astrology while thinking that it is fantasy (because of financial incentives, for example, as many great astronomers in history worked also in astrology).[5] Or someone, like Richard Dawkins, might comprehend the reasonings and alleged justifications behind creationism while being committed to evolutionary theory. Even if a theory is metaphysically unintelligible, one may be able to use it, and this usability of a theory, we take it, is an integral part of intelligibility. Therefore, we need to disentangle the scientist's metaphysical

[4] Hasok Chang makes a similar distinction and argues for a similar relationship between metaphysics and scientific practice. He writes, "Ontological principles are the basis of intelligibility in any account of reality; the denial of an ontological principle strikes one as more nonsensical than false" (Chang, 2001, p. 11). (This quote will become important for us in section 4.3) Later, he writes, "Performability requires a certain harmony within the activity. For example, any statement made within it needs to conform to the ontological principle associated with it." (Chang, 2009, p. 75).

[5] We use astrology as an example of a theory that can be scientifically intelligible without being metaphysically intelligible. In addition, this theory is not empirically adequate, but this requirement is independent of intelligibility in de Regt's framework. We also thank an anonymous reviewer for pointing out that one may have understanding of the theory of astrology without having understanding of the phenomena that astrology is supposed to explain. In other words, one can have an intelligible theory of astrological practice which will lead to understanding of astrology as a practice.

commitments from intelligibility and embed them into a wider network of one's noetic system, which comprises the scientist's background beliefs (we discuss all this in more detail in section 4.3).

Moreover, we think there is an aspect of a theory's intelligibility (and of understanding) that is not sufficiently appreciated in de Regt's account. Understanding a theory is not only a matter of being able to use it and to apply it to the phenomena. It is also a matter of being aware of what the world looks like, according to the theory, of having access to the theory's truth conditions, of knowing which kind of ontology is associated with the theory or postulated by it. A theory with an open interpretation problem, for instance, or with unclear ontological commitments, is a theory that we do not yet understand (fully), even if we are extremely successful in its application. A scientist may successfully apply a theory and have correct intuitions about qualitative results of the theory, but given that she has no clear picture of the ontology postulated by the theory, she would still miss potentially important information about the goings-on in the world. This information may not be particularly important for applying the theory in the particular context and for the particular purposes of the scientists, but it would still expose something about the world and would actually deepen the overall understanding of the scientist. The information we refer to is, for example, about what something is made of: the ontology of a phenomenon or object, or, in other words, the mechanism.

Let us go into some more details on why we think mechanisms deepen contextual understanding. What is the starting point and the aim of scientific inquiry in the first place? Grimm (2008) rightly answers that scientific inquiry starts with and aims at answering *why* something happens the way it happened and not otherwise. This answering-why has been the focus of almost the entire literature on explanation from the deductive-nomological model (Hempel & Oppenheim, 1948) to modern causal theories (Woodward, 2003). And we see this focus also in both of de Regt's theories of understanding, as well as in the theory of understanding developed by Khalifa (2017), where the main use of a theory is to explain *why* some phenomenon happens. The requirements on a theory to be intelligible and successfully usable are mainly to let the scientist efficiently use the theory. De Regt's theory of understanding, as well as Grimm's theory of scientific inquiry and much of the literature on scientific explanation, does not appreciate that apart from asking *why* something happens science also asks *what* something is made of (Hubert, 2021). Contrary to most of his contemporaries, Wesley Salmon pointed to this complementary part of scientific understanding:

> We want to know *how things work* and, it should be added, *what they are made of*. [...] What we want to do is open up the black box and see how it works. (Salmon, 1998, p. 87)

This passage bears two insights: First, we can ask what something is made of independently of why something happens. Second, we can use this information to discover *how* something works. The *how* is indeed a combination of *what* and *why:* it explains *why* something happens by means of *what* it is made of. As Salmon points out, these kinds of explanations are causal-mechanical explanations, which have recently experienced a renaissance (Glennan, 2017; Glennan & Illari, 2018; Machamer et al., 2000). Glennan (2017, p. 17) proposes the following minimal characterization of mechanisms: *A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized to be responsible for the phenomenon.* Glennan specifies very general requirements on the entities that make up the mechanism to apply this definition to a wide range of examples. What is important for our argument is that we can inquire into the entities of a mechanism for basically all kinds of phenomena. And knowing about these entities give us a deeper understanding of the phenomenon, because we would be able to answer not only "Why?" but also "What?".

In de Regt's theory, mechanical explanations are not distinguished from other types of explanations. He does not say that there is no place for mechanical explanations in the contextual theory or that mechanical explanations do not yield understanding; rather , mechanical explanations do lead to understanding (if they are intelligible, etc.), but they provide the same kind or level of understanding as other (intelligible, etc.) explanations. If a scientist uses mechanical explanations in one case and uses another type of explanation in another case, the contextual theory says that in both cases the scientist has *the same level* of understanding if the scientist is successful. As we have argued, mechanical explanations do answer two (instead of one) of the most fundamental questions we can ask about a phenomenon, and this difference, we think, needs to be incorporated into a theory of ideal understanding.

3.2 Wilkenfeld's MUDy Understanding

We think that de Regt is correct in pointing out that a theory's intelligibility plays a crucial role in understanding phenomena, but as we discussed in the previous subsection, de Regt leaves out important aspects of understanding, namely, understanding what the world is made of. Wilkenfeld (2017) criticizes de Regt's theory along similar lines and claims that de Regt ignores the significance of the *representational accuracy* of the state of understanding (which is mediated in de Regt's account by a scientific theory).[6] Wilkenfeld describes representational accuracy in the following way:

[6] Other theories of understanding that focus on representational accuracy are the following: Dellsén's dependency model theory of understanding is a theory of representational accuracy of the dependency relations

> We do not have an account of what it means for a representation to be accurate but presumably the general idea is that the actual state of affairs of the world is in some important sense similar to the state of the world as depicted in the representation. Importantly, if we assume a correspondence theory of truth, any true propositions will be representational [*sic*] accurate. (Wilkenfeld, 2017, p. 1275)

Khalifa's explanationist theory is a theory of representational accuracy as the explanatory nexus represents the explanatory relations of the world; mechanistic explanation, as we described them, represent not only causal relations but also represent what the world is made of.[7] On the other hand, theories of understanding based on representational accuracy, such as Khalifa's, tend to underestimate intelligibility. Therefore, Wilkenfeld (2017, p. 1276) proposes a *multidimensional account of understanding*, where intelligibility and representational accuracy form two dimensions:

> **Multiple Understanding Dimensions (MUD):** the quality of a state of understanding is evaluable along multiple orthogonal dimensions, including (but perhaps not limited to) both representational accuracy and intelligibility.

Wilkenfeld leaves it open whether there may be additional dimensions. If we aim at ideal understanding, we think that the two dimensions identified by Wilkenfeld need to be supplemented by (at least) three additional dimensions:

1. Truth,
2. Reasonable endorsement,
3. Fitting into the noetic system (of the scientist)

We will discuss later the details of these three additional dimensions. Regarding the truth dimension, we will call it "sufficient truth" and show how it differs from Khalifa's "approximately true" and Wilkenfeld's notion of truth. We will also show how the truth dimension relates to the other dimensions. For now, it suffices to recognize that we extend MUD by three other dimensions.

We propose a theory of ideal understanding that comprises five dimensions of understanding evaluation.

> **Ideal Understanding**: An agent S ideally understands a phenomenon P by means of a theory T, only if (i) T accurately represents P, (ii) T is intelligible

in the world (Dellsén, 2020). Le Bihan's modal account of understanding is also based on accurately representing aspects of the world, namely, the modal relations (Le Bihan, 2017).

to S, (iii) T is sufficiently true of P, (iv) S has reasonable grounds to endorse T, and (v) T fits into the noetic system of S.

We take these five dimensions to be necessary conditions for ideal understanding. It should be immediately clear that we both supplement and digress from Wilkenfeld's theory. We supplement it by three additional dimensions, but we characterize representational accuracy slightly differently than he does. Whereas Wilkenfeld is interested in the representational accuracy of the *state* of understanding, we apply representational accuracy to a scientific theory, which we regard as the most comprehensive mediator for scientific understanding, especially when we search for an ideal. One may argue that our dimension of truth (as discussed in section 4.1) coincides more with Wilkenfeld's representational accuracy of the *state* of understanding, as he seems to leave out representational accuracy *of scientific theories*.[8] On the other hand, we interpret Wilkenfeld to be open to different representational devices that yield representational accuracy of the state of understanding. In any case, we split Wilkenfeld's first dimension of representational accuracy into two: (i) representational accuracy of a scientific theory and (ii) sufficient truth based on the theory.

We are open to the possibility of more dimensions for evaluating ideal understanding; in section 5, we discuss three candidates for further dimensions (grasping, novel predictions, and transparency), but we conclude that they are inadequate for being included as necessary conditions for ideal understanding.[9]

4 Towards Ideal Understanding

In what follows, we propose three additional criteria that must be fulfilled for a theory T about a phenomenon P to succeed in providing a scientist S with the highest conceivable degree of understanding of this phenomenon.

First, T must be *sufficiently true*[10] of P. We specify what "sufficiently true" means, in a way that does justice to the role that idealizations play in understanding phenomena scientifically (section 4.1). The second requirement we identified is that for S to ideally

[8] We thank an anonymous reviewer for this insight.

[9] One may disagree with us (and with Kelp and Khalifa) that there is something like ideal understanding. In this case, you may take our account as an extension of Wilkenfeld's two dimensions of understanding, which may be named Deep MUD: the quality of a state of understanding is evaluable along multiple orthogonal dimensions, including (but perhaps not limited to) representational accuracy, intelligibility, truth, reasonable endorsement, and fitting into the noetic system.

[10] Elgin (2017) introduced a similar notion: *true enough*. We settled for a different term because our notion, although similar to Elgin's, digresses in important respects (see section 4.1).

understand P via T, S must not only endorse T; S must have *excellent epistemic reasons* to do so. In other words, S must be epistemically justified in endorsing T. We spell out what exactly this justification condition requires, and how it relates to a theory's empirical success (section 4.2). Our last criterion concerns how T relates to the already established intellectual background of the scientist. For T to provide S with ideal understanding of P, T must "make sense" relative to or *fit* to the best possible extent *into* the scientist's already established intellectual background. We explain what "fitting into" exactly means and illustrate our idea with two examples from the history of science (section 4.3).

In what follows, we explain what it takes to fulfill these three additional criteria. All the criteria we analyze can be fulfilled to a greater or to a lesser extent. Roughly, our view is that ideal understanding requires that (at least) these criteria are fulfilled to the highest possible extent. Degrees of understanding, on the other hand, will be achieved by fulfilling these criteria to a lesser extent than required for ideal understanding. Yet, we leave the detailed exploration of degrees of understanding for another project.

4.1 Sufficient Truth

Kelp (2015) demands that a scientific theory leading to maximal understanding will provide scientists with *knowledge* about the phenomenon. As knowledge requires truth (I cannot know that p if p is false), it is natural to require that the final theory about P will be true as well. In other words, it seems natural to require that the dependence relations postulated by the theory will correspond to real dependence relations (Dellsén, 2020); that the claims the theory makes about possible worlds will be correct; that the ontology postulated by the theory will correspond to the real ontology, and so on. Let us call this the *strong factivity constraint* for maximal understanding:

Strong factivity constraint

For a theory T about P to provide a scientist with maximal understanding of P, T must contain only truths about P.

And yet here is an argument that non-factivists might raise against this constraint (Elgin, 2017). Our current best science is full of "falsehoods". More precisely, it is full of representational systems that are known to *misrepresent* the way the world actually is. A good example is an idealized model (Cartwright, 1983; Giere, 2004; Morgan & Morrison, 1999; Potochnik, 2017). The ideal gas law, for instance, accounts for the behavior of (real) gases by describing the behavior of an *ideal* gas comprised of molecules that are dimensionless or perfectly spherical and exhibiting no intermolecular forces. Clearly, there is no such thing in the real world; nevertheless, our best scientific theories may still be (even only

approximately) on the right track. [11] By depicting gases *as if* they had those features, however, the ideal gas law enables scientists to predict and to understand how real gases behave. Certain falsehoods, thus, seem to figure centrally in our scientific understanding of the real world (Elgin, 2017, p. 61).

Moreover, Elgin claims, these falsehoods are not simply "necessary evils" in a world that is too complicated for us to deal with; they are not simply tools that scientists use to approach the true description of reality:

> Idealization is not taken by scientists to be an unfortunate expedient, but rather to be a powerful tool. Although they expect today's idealizations to be replaced, they harbor no expectation that in the fullness of time idealizations will be eliminated from scientific theories. … Elimination of idealization is not a desideratum. Nor is consigning them to the periphery of the theory. … The ideal gas law lies at the core of thermodynamics, and some such model is likely to lie at the core of any successor to current theories. (*Ibid*. p. 62)

If Elgin is right about this, and the elimination of idealizations from scientific theories is not a desideratum, our strong factivity constraint for full understanding is too demanding. Even when the Theory of Everything is found, if it exists in the first place, idealizations will remain a crucial part of scientific practice.

There are at least two ways to answer this objection. First, one could argue against Elgin and claim that whatever role idealizations play in science, the same role can be played at least as effectively by representational systems containing nothing but the truth (Le Bihan, 2019; Sullivan & Khalifa, 2019). If this were the case, we would have no reason to believe that the final theory providing ideal understanding will contain idealizations. Idealizations, contra Elgin, would turn out to be dispensable. If this line of argument works, then the strong factivity constraint holds.

Another way to answer would be to grant Elgin's claim that certain "falsehoods" such as idealizations will figure centrally even in the final edifice of science, and still defend that full understanding *as an epistemic state* requires truth. This is not as paradoxical as it might sound.

[11] Khalifa (2017) uses "being on the right track" in a different sense than we do. For him, agents (not theories) can "be on the right track" to having explanatory understanding. He writes on p. 86, "In this case, to claim, for instance, that 'Tom understands ecology' is simply shorthand for, e.g., 'For some p about ecology, Tom is on the right track to understanding how/why p.' Here, 'being on the right track' means that one does not know the proper answer to a relevant explanation-seeking question, but one has information that would be useful in acquiring such knowledge."

In spelling out the relation between understanding and truth, there are two levels of correctness to distinguish. One is the level of the correctness *of the representational systems* that scientists use; the other is the level of the correctness of the information that scientists *believe* based on the representational systems in question. The two levels are certainly related, but they are independent. Even a *false* representational system could in principle work as a source of accurate/true information about reality— given that one is equipped with the right background knowledge. Take for example an idealized model. An agent who properly understands the model will be aware that the model is idealized. In the best-case scenario, she will be aware of which information contained in the model is to be interpreted as a realistic representation of its subject matter and which, instead, is merely fictional— and she will believe those parts of the model that correspond to reality and merely accept or disregard the fictional parts. But given that one has reached this vantage point, everything one will *believe* about reality on the basis of the model will be true after all, no matter how well and how accurately the model actually represents its subject matter (Greco, 2014; Lawler, 2019; Mizrahi, 2012; Nawar, 2019; Strevens, 2008).

So, even if Elgin were right in claiming that certain "falsehoods" such as idealizations will be involved in the final theory T about P, we suggest, scientists who will understand P fully on the basis of T will have the appropriate *skills* necessary to identify and extract the true information contained in these falsehoods. Hence, everything scientists will believe about P and about P's subject matter *on the basis* of T will be true. So, we argue in order to be closer to actual and realistic scientific practice, the following *weak factivity constraint* will certainly hold:

Weak factivity constraint

For a theory T about P to provide a scientist S with maximal understanding of P,
everything S *believes* about P on the basis of T must be true.

De Regt and Gijsbers controversially claimed that understanding should be conceived as a non-factive cognitive state, *because* "understanding can be gained from representational devices that are false, and not just slightly false, but wildly so" (De Regt & Gijsbers, 2017, p. 50). We are now in the position to see that this is a *non sequitur*. Even if the representational systems we use are wildly incorrect, we might have reached the vantage point and developed the skills necessary to identify and to extract the true information contained in them. But if we have, then the understanding we will gain based on these inaccurate representational systems will be factive after all; that is, everything we will believe based on the inaccurate representational systems we use will be true. This may be still hard to accomplish in practice, especially synchronically without the wisdom of hindsight, but the weak facticity constraint is more realistic than the strong one. At least in model building, scientists are familiar with the

problem of deriving truths from idealizations, that is, from intentional falsehoods. The caveats of extracting truths from false theories have been debated between Hasok Chang (2003) and Stathis Psillos (1999) on the plausibility of preservative realism, where Chang emphasizes the problems with such an endeavor.[12] We want to discuss a more optimistic example in the following.

De Regt & Gijsbers (2017) discuss the phlogiston theory as an example of a theory that, despite being "wildly incorrect", succeeded in providing scientists with an understanding of phenomena. Pace de Regt & Gijsbers, phlogiston theory is in fact a great example of a theory that can work as to generate (approximately) *true beliefs* for a scientist who is equipped with the right background knowledge (more precisely: for a scientist who masters modern chemistry). Phlogiston theory explained the process of combustion roughly in this way: when a substance burns, an unobservable matter called "phlogiston" *leaves* the substance in question— usually in the form of a hot flame or evaporating gas. This has long been considered as "wildly incorrect", because, according to oxygen theory (the theory that superseded phlogiston theory), when a substance burns, nothing is lost, but something (oxygen) is *added* to it (this process is called oxidation). Modern chemistry, however, uncovered the mechanism underlying oxidation and made us realize that phlogiston theory was actually not as incorrect as it might seem: in an oxidation process, a chemical bond forms between an electropositive substance (such as metal, coal) and an electronegative substance (such as oxygen). In this chemical bond, the electropositive substance *gives up* some of its electrons; more precisely, the electropositive substance *donates* some of its electrons to the electronegative one. Modern chemistry, thus, enables us to see that phlogiston theory was right, or at least on the right track, in claiming that something in the process of combustion "gets lost".[13]

We call our notion of truth, based on the weak factivity constraint, *sufficient truth*. We want to emphasize the pragmatic aspect of truth seeking in scientific practice, which is part of understanding. It is rarely the case that one can read off truth of a theory. Even if the theory that we use is representationally accurate, it is often not possible to directly apply the theory. Idealizations and model building are necessary paths in "applying" the theory. Therefore, we added the truth dimension to representational accuracy. On the other hand, if a theory is representationally inaccurate, experienced scientists can still extract truths from it.

[12] We thank an anonymous reviewer for this reference.

[13] Schurz (2009, p. 109) puts it in terms of a correspondence relation holding between phlogiston theory and modern chemistry: the loss of phlogiston in phlogiston theory *corresponds* to the donation of electrons from an electropositive substance to an electronegative one that, according to modern chemistry, takes place in the process of oxidation. (For further attempts to reconstruct phlogiston theory as an approximately or partially true theory, see Falguera & de Donato Rodríguez, 2016; Ladyman, 2011).

Newtonian mechanics is a good example because we know that it is not the true theory of the universe, but we still use it successfully in specific domains (for example, when sending a rocket to the moon).

Wilkenfeld (2017) starts with a correspondence theory of truth, when he writes, "Importantly, if we assume a correspondence theory of truth, any true propositions will be representational [*sic*] accurate." (p. 1275). The correspondence theory of truth is in principle adequate, and the way Wilkenfeld formulates it, it seems plausible: true propositions correspond to facts in the world and are representationally accurate. In practice, however, the correspondence theory of truth is either useless or misleading, because " approximately true propositions and non-propositional representations (e.g., maps) can be more or less accurate as well" as Wilkenfeld acknowledges (p. 1275). Therefore, he works with a generalized correspondence theory to accommodate these situations. Still, the practical aspect of extracting truth from representationally accurate and even inaccurate systems is not sufficiently appreciated in Wilkenfeld's MUDy understanding.

Khalifa (2017) also uses a notion of truth that differs from ours. He eventually distinguishes between *strict truth* and *approximate truth*, but he locates the pragmatism of truth seeking not in the notion of truth itself, which is also based on the correspondence theory, but in his notion of explanation and more explicitly in his notion of knowledge.

Recall that Khalifa uses explanation in the following sense:

Q (correctly) explains why P if and only if:

(1) Q is (approximately) true;
(2) Q makes a difference to P;
(3) Q satisfies your ontological requirements (so long as they are reasonable); and
(4) Q satisfies the appropriate local constraints.

That proposition Q is said to be (approximately) true can be understood, as Khalifa emphasizes, in two different ways. In the realist sense, the proposition Q is true means that Q is representationally accurate, as in Wilkenfeld's case; that is, Q corresponds to facts in the world. If Q is approximately true, Q is close to be representationally accurate. In the anti-realist sense, the proposition Q is true means something weaker, for example, being empirically adequate.

Now, Khalifa defines when explanations are approximately true:

> Let's say that *"[Q] explains why [P]"* is approximately true if and only if the first of these conditions is satisfied, and, furthermore, *some* of the terms in the explanans *[Q]* that purport to make a difference to the explanandum *[P]* actually do make a difference and also satisfy your preferred ontological requirements. (Khalifa, 2017, p. 55)

An explanation is approximately true, if Q (the explanans) contains some, but not all, difference-makers of P (the explanandum). If the explanation contained all difference-makers,

it would be *strictly true*. Because in practice, we cannot and often do not want to identify all difference makers, approximately true explanations are true enough for all practical purposes. How can scientists know that an explanation is the right one? Here, Khalifa explicitly mentions a specific truth-seeking process:

> S has scientific knowledge that *[Q] explains why [P]* if and only if the safety of S's belief that *[Q] explains why [P]* is because of her scientific explanatory evaluation (SEEing). (Khalifa, 2017, p. 12)

The scientist S needs to evaluate the particular explanation, and this evaluation justifies the belief of the scientist that the explanation is correct. This justification rules out the case that the agent forms a true belief due to coincidence; thus, the belief is said to be safe. The evaluation process, called *SEEing*, has three parts:

1. consideration,
2. comparison (of the potential explanations),
3. belief-formation.

First, the scientist needs to *consider* other potential explanations for the phenomenon P. Second, the scientist needs to *compare* all these explanations according to best scientific practice. After that, the scientist forms a belief about the different potential explanations, and in the best case singles out an explanation as the best or true one. These three steps comprise the pragmatic procedure for scientists to extract truth from many candidate explanations similar to the weak factivity constraint. Khalifa, however, restricts the SEEing process to explanations, whereas we are open to different representational devices, such as models, diagrams, etc. based on a theory that are used to extract sufficient truth.

4.2 Reasonable Endorsement

We have been working with the assumption in the background that for S to understand P via T, S must entertain a doxastic or noetic attitude of some sort towards T. We briefly mentioned that S must *endorse* T or *commit* herself to T[14] . We have not yet said anything, however, about the *normativity* of such endorsement involved in understanding. When is a

[14] There is no agreement in the literature concerning the exact nature of the commitment involved in understanding. Some scholars take it to be a form of belief; others take it to be a form of acceptance. We do not take a stance on this issue, but we do want to leave open the possibility that ideal understanding will be embodied in and generated by representational systems that comprise non-propositional parts. This is also the main reason why we do not exclude the possibility that while ideal understanding involves knowledge, it cannot be reduced to knowledge. However, we leave the exploration of this idea for another project.

scientist's endorsement of a theory *justified*, or *warranted*? When is it *rational* for a scientist to endorse a theory?

Many authors point out that for a subject to be justified in endorsing a theory (or an explanation), such theory must turn out to be the best of (or "the winner" among) all available alternatives. Dellsén's optimality model of understanding (Dellsén, 2019), Elgin's reflective equilibrium model (Elgin, 2017), and Khalifa's SEEing model (Khalifa, 2017), e.g., rest on this idea. It should be noted, however, that this normative principle admits of (at least) two readings, depending on how one chooses to spell out the phrase "the best of all available alternatives". On what we might call the *subjective* reading, for a subject to be justified in endorsing a theory T about P, T must turn out to be the best of all the theories about P that the subject took into consideration (and ruled out on evidence-based grounds). On the *objective* reading, on the other hand, for a subject to be justified in endorsing a theory T about a phenomenon P, T must turn out to be the best of all the alternatives available *in the epistemic environment* of the subject – regardless of whether the subject actually considered these alternatives or not.

To better grasp the difference between these two readings, imagine the following scenario. It is January 2020. A GP is trying to understand the condition of a patient showing light flu symptoms: cough, sore throat, runny nose. The GP does everything in her ken to make the correct diagnosis (e.g.: she considers the medical history of the patient with great care; she rules out pneumonia after finding out that the patient has no fever; she rules out bronchitis after checking the condition of the patient's lungs; and so on). The result of her assessment is that the patient has a common seasonal cold she should not really worry about. The diagnosis is correct. However, the very same symptoms could have very easily signalized a much more dangerous health problem: a COVID-19 infection. The WHO sent an emergency alert about this possibility to all registered physicians a couple of hours before the GP made her diagnosis, but the GP did not have access to her email account because of a problem with her Wi-Fi. Now, was the GP justified in assuming that her patient had a seasonal cold (before reading the WHO alert)? The answer here is probably both yes and no. Subjectively, i.e., given her best effort and her best judgment, probably yes; objectively, i.e., given the objective probability that her diagnosis was the correct one, probably no. The seasonal cold was the best alternative among the alternatives that she considered, but not among all alternatives available in her epistemic environment.

We suggest that in the perfect epistemic situation of ideal understanding that we are describing, the subjective will match the objective; that is: the scientist's subjective confidence that the theory accounting for P is true will match the theory's real, objective confirmation. In other words: a scientist understanding a phenomenon P via a theory T (about P) to the highest conceivable extent will be both subjectively and objectively justified in endorsing T.

Within De Regt's theory of understanding, a scientific theory must be reliably empirically successful to provide a scientist with understanding (De Regt and Gijsbers 2017: 55; see also De Regt 2017: 119). We agree with De Regt on this point. We take this requirement to hold for understanding generally, and to hold *a fortiori* for ideal understanding. We demand from an "understander" of P to be able to successfully predict P's occurrence and P's future development. (We probably would not say that one understands the phenomenon of climate change if she predicted that temperatures on earth will decrease soon, while instead they end up increasing.) Moreover, we demand from an understander of P to be able to successfully manipulate P's domain as to influence in some way P's occurrence or its development. (We probably would not say that a scientist understands the spread of the COVID-19 if she devised measures meant to reduce the spread of the virus that end up fostering it instead.) However, we want to point out that the reason why a theory's empirical success is relevant for understanding is that a theory's empirical success provides scientists with good (despite *prima facie* and defeasible) *epistemic reasons* to endorse the theory. In other words, a theory's empirical success contributes to the scientist's justification in endorsing the theory. [15]

4.3 Fitting into the Noetic System of the Scientist

"Ontological principles are the basis of intelligibility in any account of reality; the denial of an ontological principle strikes one as more nonsensical than false," writes Chang (2001), and we want to take the core of this idea to develop our final necessary criterion of ideal understanding: the theory T must also "make sense" relative to or *fit* to the best possible extent *into* the scientist's already established noetic system.

A couple of clarifications are in order here. First, we suggest calling "noetic system" the set of informational units that a particular subject believes, accepts, or endorses at a certain moment in time.[16] Some of these informational units will be true, some will be false; some will

[15] Some authors, among them Dellsén (2017) and Hills (2016), have argued that understanding does not require justification (at least not in the same way in which knowledge does), as one allegedly might understand why p, or that p is because of q, despite having a defeater for q. Nothing we say in this paper rules out the possibility that the threshold of justification required for low degrees of understanding might be lower than the threshold required for knowledge. However, we do take it to be plausible that a no-defeater-condition will apply at least for understanding in its highest conceivable form.

[16] Noetic comes from the Greek word *noētikos* meaning "intellectual," which is derived from *noein:* to think (from *nous:* mind). We chose to talk of a scientist´s noetic, and not doxastic system, because we want to take into account not only the informational units that a scientist takes to be true, but also those that she accepts, e.g., for practical or reasoning purposes. Bengson also uses the term "noetic" to characterize his theory of understanding. His motivation is similar to ours: he chose the word noetic because it refers to thinking or the intellect in general and is not limited to explanation and belief (Bengson 2015: 3).

be close enough to the truth to serve a particular epistemic, cognitive, or practical aim. The informational units belonging to a noetic system will not be isolated, i.e., they will not form a long conjunction. Rather, they will depend upon one another in many ways (logically, semantically, explanatorily, probabilistically, epistemically, and so on). A noetic system, thus, can be represented as a structured set, and it shows two dimensions: an *informational* and a *relational* one. (see Malfatti, 2021; Schurz & Lambert, 1994)

But what does "fitting" exactly require? What does it take for a theory to fit into a scientist's noetic system? Fitting, as we conceive it, involves a *negative* and a *positive* requirement. The negative requirement amounts to the following: how well a theory T fits into a noetic system W is inversely proportional to the number of contradictions and/or cognitive dissonances arising from the conjunction between T and W (or from the incorporation of T in W). So, the more contradictions and/or cognitive dissonances generated by the conjunction of T and W, the poorer the fitting of T into W. Absence of contradictions and/or cognitive dissonances is not enough for fitting, however (as Bartelborth, 1999, among others, has argued). Suppose a scientist endorses a theory T that is *isolated* in her noetic system, and that shows very little positive connection at all with other theories "in the neighborhood" that bear on it. We would probably hardly say that the theory "makes sense" relative to or fits well into the scientist's noetic system. So, we suggest that how well a theory T fits into a noetic system W will be directly proportional to the number (and strength) of connections bounding T to other theories in its neighborhood and that are relevant to it.

We will not make heavy weather of the claim that, at least typically, resolving contradictions, ironing out cognitive dissonances and enhancing the systematicity of one's noetic system improve one's epistemic standing and advance one's understanding of reality. Of course, specifying exactly which theories and assumptions "in the neighborhood" or "in the background" are relevant for a certain theory (and for understanding phenomena via the theory) will not always be a straightforward matter. It is quite straightforward, however, to assume that whatever scientific theory we formulate for empirical phenomena cannot contradict and must be somehow (explanatorily) connected to our "immanent picture" of the world. In other words, such a theory cannot contradict the deliverances of our perception, and in case it seems to do so, it must have the resources to explain why humans perceive things as they do, given that at a fundamental level, reality is very different from how it appears. It is also plausible that whatever scientific theory we reasonably endorse will be "in equilibrium" (and ideally also properly connected) with our more general metaphysical assumptions about reality. These metaphysical assumptions are partially derived from everyday experiences, but sometimes derive from the historical development of science.

To substantiate our claims, we will now discuss two historical cases in physics that show that a theory needs to fit into a scientist's noetic system to provide her with understanding. We take this requirement to hold for understanding generally, and to hold *a fortiori* for ideal understanding. Of course, fitting is not an all-or-nothing matter; it admits of degrees, as understanding does. So, we suggest that for a theory to provide a scientist with the highest conceivable degree of understanding, the theory must fit *to the greatest possible* extent into the scientist's noetic system. Lesser degrees of understanding, on the other hand, will be compatible with lesser degrees of fitting.

4.1. Example 1: Newton and Action at Distance

When Newton published his *Principia* in 1687, the received worldview was Descartes' corpuscularist physics as described in *The World* and in the *Principles of Philosophy* that required that physical bodies need to act by contact. Therefore, Newton's contemporaries had a hard time accepting the possibility of material bodies acting upon each other without an intervening medium. Even Newton himself struggled with this aspect of his theory of gravitation (de Regt, 2017, p. 116; Henry, 2017), when he expressed his famous "Hypotheses non fingo." (Newton, 1999, p. 943). What he offers in the *Principia* is a valid law of gravitation that contains all and not more than the information he could extract from observation. In particular, the law does not postulate a mechanism for gravity. Since the observations during Newton's time do not distinguish any of the possible mechanisms, Newton suspended judgment. This is also reflected in the title of his work. In contrast to Descartes, who described in detail the mechanisms for his physics in his *Principles of Philosophy,* Newton chose *Mathematical Principles of Natural Philosophy,* that points to only one aspect of a complete theory of physics, namely, the necessary mathematics for empirical predictions.

As de Regt, 2017 (p. 117) puts it: "The notion of action at a distance flatly contradicted the principle of contact action and was therefore unacceptable as an explanatory resource." The reluctance of Newton's contemporaries and of Newton himself to fully endorse his theory as complete, since it contradicted already endorsed and established principles, suggests that agents *typically* strive for achieving consistency in their noetic systems when they try to make sense of reality. And yet it could be argued, first, that this is the way it should be: it was *rational* for Newton and his contemporaries not to (fully) endorse such a theory before having revised their metaphysical commitments of supporting contact action. And second, it is hard to deny that those scientists who had changed their metaphysical worldview and revised their corspuscularist commitments would have understood phenomena *better* according to Newton's theory than those scientists who had decided to stick to Cartesian metaphysics (if Newton's theory had turned out to be true).

It should be noted here that Newton and his rivals, Leibniz and Huygens, were experts in applying and working with Newtonian mechanics, even if this theory clashed with their metaphysical background beliefs and principles. Therefore, Newtonian mechanics was an intelligible theory (in our conception of intelligibility), but Newtonian mechanics led to dissonances in the noetic system of scientists who followed Cartesian metaphysics.

De Regt takes a slightly different morale from this historical case. For him, there is another kind of intelligibility that we need to consider:

> The type of intelligibility they demanded from a theory may be called *metaphysical intelligibility*, where a theory is metaphysically intelligible if it harmonizes with extant, or preferred, metaphysics. (de Regt, 2017, p. 160)

De Regt's original definition of intelligibility, which he sometimes refers to as scientific intelligibility, is about how well a scientist can use and apply a scientific theory. Metaphysical intelligibility, on the other hand, is not about how well someone can use and apply the theory, but rather how well the theory matches with metaphysical or ontological principles and frameworks. As we argued in 3.1, we think it is a bit odd to put these two concepts under the same umbrella and call them two different forms of intelligibility.[17] It is not merely a choice of words to dub the consistency with metaphysics a form of intelligibility, because different contexts may demand consistency with other parts of one's noetic system than beliefs about metaphysics, like, for example, the immanent picture of the world, empirical evidence, meta-laws (like Bell's theorem or the laws of thermodynamics), methodological principles, etc. Embedding metaphysical theories within the noetic system helps us to recognize that there can be more dissonances than between a scientific theory and its metaphysics.

4.2. Example 2: Einstein and Quantum Non-Locality

The principle of locality has it that the causal influence on an object can be exercised only by other objects located in its immediate surroundings. This means that if an object *a* has a causal influence on an object *b* which is *not* located in *a*'s immediate surroundings, then there must be an object *c* mediating between *a* and *b* and carrying the causal influence from *a* to *b*. A theory is called "local" if it incorporates and respects the principle of locality.[18] According

[17] De Regt (2017, p. 160) is aware of this concern when he writes in a footnote, "Note that metaphysical intelligibility differs from intelligibility as defined in section 2.3 [of his book]. However, as will be argued below, there can be overlap and interaction between the two."

[18] Depending on the particular application, different concrete definitions of locality have been proposed (for instance, Belot, 1998; Lange, 2002; Maudlin, 2014).

to some interpretations, quantum theory is not a local theory. Some quantum phenomena (those due to entanglement) seem to violate the principle of locality. Microscopic objects far apart from each other seem to be able to "communicate" or influence each other *simultaneously*.

Einstein famously struggled with this aspect of quantum theory and found it very hard, if not impossible, to accept (see, for example, the EPR thought experiment in Einstein et al., 1935; and the historical discussion in Becker, 2018). What grounded Einstein's skepticism clearly was a worry of overall coherence of his noetic system. Accepting quantum theory in a non-local interpretation would have generated clear tensions with other (classical) theories that he already endorsed and that he was not willing to give up: classical electrodynamics, on the one hand, and his own theories of relativity, on the other. We take it to be undeniable that Einstein would have been in a better epistemic position and would have understood phenomena better (based on quantum theory, based on his theories of relativity, or based on a completely different theory) if the tension in his noetic system would have been resolved. Einstein certainly had a very deep understanding of quantum mechanics; therefore, he was able to see the tension with locality. We even think that quantum mechanics was intelligible for Einstein because he had sufficient intuitive grasp of this theory and had the skills to apply it. We claim, however, that Einstein didn't have *ideal* understanding of quantum mechanics, because quantum mechanics was incoherent with other physical theories and with what he believed physics to be. Probably, Einstein would not say, "I don't understand quantum mechanics!", but rather "Quantum mechanics does not make sense!" This cognitive dissonance may not be relevant for most applications, as we can see in the history of quantum mechanics and the current way physicists use and teach quantum mechanics. Many do agree that quantum mechanics in its textbook form doesn't make real sense, but it is so successful in making predictions and building technology that they simply live with this dissonance.

Suppose that Einstein was right, and quantum mechanics has to be a local theory.[19] One might worry that, according to our account of ideal understanding, Einstein would have understood quantum mechanics better than his rivals, who did not see a dissonance between quantum mechanics and their background metaphysical beliefs.[20] A detailed answer would require a theory of what it means to improve understanding, which we do not pursue in this paper, but we see different possibilities to reply to this objection. First, it is reasonable to say that Einstein, in a certain way, understood quantum mechanics better than some of his colleagues, because he pointed to shortcomings of the theory that others did not acknowledge. Even if he scored worse along the dimension of fitting than the others, he overall understood quantum mechanics better. Second, it is also reasonable to say that Einstein did not understand

[19] Indeed, quantum mechanics can be made local and still be empirically adequate with superdeterministic or retrocausal mechanisms, but these subtleties are not relevant of the example we discuss.

[20] We thank an anonymous referee for raising this concern.

quantum mechanics better than his rivals, but that his background beliefs were better or more adequate (since we suppose that the world was indeed local). In any case, our model of ideal understanding allows us to recognize that contradictions and dissonances have to be taken seriously, because they signalize a failure in understanding that ought to be resolved somehow. The mere absence of dissonances, of course, is no guarantee that understanding succeeds—this is also why we added, among other things, a factivity requirement.

5 Unsatisfactory Criteria for Ideal Understanding

Apart from our five conditions, one may demand further conditions for ideal or deep understanding. Grasping, novel predictions, and transparency are prima facie plausible candidates. We argue in the following that these conditions do not help us in understanding phenomena better or even ideally.

5.1 Grasping

The talk of "grasping" has become a commonplace in the literature on understanding. Prominent authors argue that whatever understanding is, it crucially involves grasping[21]. Given that this is supposed to hold for degrees of understanding and for understanding generally, it might seem *prima facie* strange or suspect that grasping has no place whatsoever in our account of ideal understanding. We think, however, that this worry is misguided.

Exactly what is grasping? What does it mean to grasp something – e.g., a phenomenon, fact or representational system? There is no agreement in the current literature on how these questions should be answered.

Some authors link grasping to the specific phenomenology of understanding. It is tempting to describe the experience of coming to understand by saying that when understanding succeeds, we "see" or "grasp" something, e.g., how things fit together. In other words, there is a way it feels like, when we come to understand— and "grasping", or so it is argued, is an effective way to describe this process from the first-person perspective. If this explication is correct, grasping might or might not be involved in ideal understanding, but this question is

[21] Bengson (2017, p. 19), e.g., in his general characterization of the profile of an understander (which he calls U-profile), takes it to be a platitude that "to genuinely understand something is to grasp it—whatever is understood—in such a way that it makes sense to you". Bengson does not offer a specific account of grasping, but he seems to take grasping to be the fundamental cognitive relation between the subject of understanding and the object being understood, along the lines of Strevens (2008).

simply uninteresting for us, as it lies outside the scope of epistemology and philosophy of science.

Other authors defend the view that the notion of grasping performs an essential explanatory role in any theory of understanding. Consider the following. We can truly believe, maybe even know, a great deal about something, and yet fail to understand it. Why? How is this possible? The notion of grasping is typically introduced to account for this fact. Understanding, or so it is argued, is not a matter of simply assenting to information or propositions. We can assent to a certain piece of information, we might even have excellent reasons to believe that it is true, while it remains completely "inert" in our noetic system. When, in contrast, we *grasp* a piece of information, or so it is argued, this activates, so to say; it becomes like an instrument that we can use - e.g., as a basis for inference, explanation, prediction, action (Grimm, 2011, 2021; Hills, 2016). Grasping, thus, seems to enable one to fill the gap between the theoretical and the practical aspect of understanding.

We think that the ability to use a certain piece of information for a variety of cognitive and practical purposes is indeed an important aspect of understanding. We do think that understanders (and of course, most prominently, ideal understanders) are not successful just theoretically, but also in the practical domain. We just do not think that one needs to postulate a mysterious mental act of grasping to do justice to this aspect of understanding. It is among other things because we felt the pressure of filling the gap between the theoretical and the practical aspect of understanding that we followed de Regt's path in introducing an intelligibility requirement for understanding. Given that we understand a certain theory or representational system, it is quite natural that we will develop a set of abilities to explain, draw inference, make predictions, maybe even act. But then, the talk of "grasping" turns out to be unnecessary and redundant, as everything epistemologically significant about grasping can be expressed by using more familiar and less controversial epistemological notions. (See also Khalifa 2017, 14 and 79, for a similar reductionist proposal.)

5.2 Novel Predictions

We have claimed that for a theory to provide scientists with ideal understanding, it must work as a reliable basis for action and prediction. In other words, it must be reliably empirically successful. We have also pointed out that the empirical success of a theory contributes to understanding somewhat indirectly, by providing scientists with good reasons to endorse the theory. But as we are concerned with the highest conceivable form of understanding, why not require more than mere reliability in accommodating known phenomena? Why not require the highest conceivable form of empirical confirmation, e.g., the one provided by novel predictions? Note that to count as "novel", a theory's prediction must be, among other things,

informative and *a priori* improbable (Alai, 2014).[22] Suppose that an astronomical theory predicts that "there is an unknown planet somewhere in the universe". Such an existential claim is so vague and scarcely informative that it is extremely likely to come out true – and this a priori, i.e., independently from our particular prior odds. If it does indeed come out true, the degree of confirmation it will provide to the theory from which it follows will thus be very low. The situation would be very different if the theory succeeded in predicting that "there is an unknown planet with mass x with orbit y in the solar system". Such a prediction would be confirming, given how informative it is and how unlikely it is to come out true.

While novel predictions are certainly crucial for scientific progress, we do not think that they play a crucial role in understanding, nor do we think that we should require from a theory providing scientists with ideal understanding to enable novel predictions. To see this, consider the following scenario.[23] Suppose that we unify quantum field theory and general relativity into an astonishingly accurate theory T with 23 fundamental constants. T enables us to make multiple novel predictions, so we are very confident that the theory is correct. Moreover, T satisfies all our other criteria for ideal understanding. Then, 1,000 years later, scientists formulate a new theory, T*, which is empirically equivalent to T, but is much simpler, as it has only 3 fundamental constants. T* will arguably make no novel prediction: it will predict exactly what we already expected. And yet, it would be very odd to say that this theory will not provide scientists with better understanding.

5.3 Transparency

Linda Zagzebski famously argues that transparency is the hallmark of understanding. Transparency, in Zagzebski's view, is also what makes understanding different from knowledge. She writes:

> Understanding has internalist conditions for success, whereas knowledge does not. Even when knowledge is defined as justified true belief and justification is construed internalistically, the truth condition for knowledge makes it fundamentally a concept whose application cannot be demonstrated from the inside. Understanding, in contrast, not only has internally accessible criteria, but it is a state that is constituted by a type of conscious transparency.

[22] Hitchcock & Sober (2004) also discuss the role of novel predictions for confirming or supporting a scientific theory. They warn that scientists may overfit their data with new theories they build. As they mention in their paper, the debate on the significance of novel predictions goes back to William Whewell (1840), for whom novel predictions are crucial for science, and John Stuart Mill (1843), who said that prediction and fitting (accommodation) do only psychologically differ.

[23] We thank Charles Sebens for this example.

> It may be possible to know without knowing that one knows, but it is impossible to understand without understanding that one understands.... [U]nderstanding is a state in which I am directly aware of the object of my understanding, and conscious transparency is a criterion for understanding. (Zagzebski, 2001, pp. 246–247)

Is ideal understanding transparent? There are actually two questions to be tackled here:

(i) When a scientist S ideally understands a phenomenon P, is S always also *aware* that this is the case?

(ii) When a scientist S *believes* to understand a phenomenon P ideally, does this mean that S also *actually* understands P in the ideal sense?

We think that both questions should be answered in the negative. It has been widely acknowledged in the literature in epistemology and philosophy of science that agents are not infallible in self-reflection. I.e., they are not infallible in their evaluation of their own cognitive states. Think of the so-called "illusions of understanding" (Rozenblit & Keil, 2002). Very often, we might think we understand, while in reality, we do not. Very often, we think to have identified a causal pattern in the real world, while actually, the only pattern we "see" is one within our *representations* of the world. This holds for understanding generally, and there is no reason to think that ideal understanding will be any different. And on the other hand, it is plausible to assume that agents might think that there is still work to be done to reach ideal understanding, while actually, the top has already been reached. This is because ideal understanding requires a very high degree of justification, but not certainty.

All things considered, then, we have reason to think that ideal understanding is not transparent. Moreover, we would like to point out that transparency might not even be *desirable* for understanding. If every time we reached understanding (ideal or non-ideal) we were also infallibly aware and certain of this, we would consider things settled and we would tend to stop inquiring. We would lose the flexibility to revise and rearrange our views. We would tend to disregard alternative perspectives and run the risk of becoming closed-minded. In doing so, we might miss opportunities for epistemic improvement and self-correction. Uncertainty, on the other hand, makes us fallibilist and epistemically humble. And as fallible and epistemically humble human beings, we are never certain that our views are correct. Therefore, we keep questioning, analyzing, testing, and challenging our theories. And it is by doing so, by being

open to the possibility that, despite our best efforts, we might be fundamentally wrong, that we foster the advancement of science.[24] As John Stuart Mill puts it, in his famous essay *On Liberty:*

> If even the Newtonian philosophy were not permitted to be questioned, mankind could not feel as complete assurance of its truth as they now do. The beliefs which we have most warrant for, have no safeguard to rest on, but a standing invitation to the whole world to prove them unfounded. (Mill, 1859, p. 23)

6 Conclusion

We have analyzed the epistemic state of ideal understanding. We have identified five necessary criteria that must be fulfilled for a theory T to succeed in providing a scientist S with ideal understanding of a phenomenon P. In particular, we have argued, (i) T accurately represents P (representational accuracy); (ii) T must be intelligible to S (intelligibility requirement), (iii) T must, at least, work as a basis of only true beliefs about P and P's subject matter (factivity requirement); (iv) S must be epistemically justified in endorsing T (rational endorsement requirement) and (v) T must fit into S's noetic system (fitting requirement). Ideal understanding might turn out to be a practically unreachable epistemic state. Nevertheless, our analysis sets a reference for how close a scientist approaches this ideal and how we may improve our current theories, on the one hand, and our overall epistemic position, on the other, to achieve a deeper understanding of the world.

Acknowledgements: We wish to thank Jeffrey Barrett, Gabriele Carcassi, Maaneli Derakhshani, Henk de Regt, Joshua Eisenthal, Christopher Hitchcock, Kareem Khalifa, JiMin Kwon, Frederick Eberhardt, Davide Romano, Charles Sebens, Michael Strevens, and the audience of the *Caltech Philosophy of Physics Reading Group* and the *3rd Scientific Understanding and Representation Workshop* for very helpful comments and discussions. We

[24] One might object here: once ideal understanding has been reached, it simply does not make any sense to pursue inquiry further! All the questions will be answered. So, it is not clear why a fallibilist stance would be desirable or advantageous for an ideal understander. It should be noted, however, that the epistemic state that we tried to characterize in this paper is the one of understanding *a particular phenomenon P* ideally. Understanding a particular phenomenon ideally, however, is very different from understanding ideally *every phenomenon that can be understood.* Our point is that by continuing to inquiry about P, even if we have already reached the state of ideal understanding of P, we might end up improving our overall understanding of reality, i.e., our understanding of other domains and other phenomena "in the neighborhood" of P. We thank an anonymous reviewer for pressing us on this point.

also wish to thank two anonymous reviewers for their incredibly helpful and constructive comments that significantly improved the paper.

7 References

Alai, M. (2014). Novel Predictions and the No-Miracle Argument. *Erkenntnis*, *79*(2), 297–326.

Bartelborth, T. (1999). Coherence and explanations. *Erkenntnis*, *50*(2–3), 209–224.

Becker, A. (2018). *What Is Real? The Unfinished Quest for the Meaning of Quantum Physics*. New York: Basic Books.

Belot, G. (1998). Understanding Electromagnetism. *The British Journal for the Philosophy of Science*, *49*(4), 531–555.

Bengson, J. (2015). A noetic theory of understanding and intuition as sense-maker. *Inquiry*, 58(7-8):633–668.

Bengson, J. (2017). The unity of understanding. In Grimm, S. R., editor, *Making Sense of the World: New Essays on the Philosophy of Understanding*, chapter 2, pages 14–53. Oxford: Oxford University Press.

Bogen, J., & Woodward, J. (1988). Saving the Phenomena. *The Philosophical Review*, *97*(3), 303–352.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford: Clarendon Press.

Chakravartty, A. (2017). *Scientific Ontology: Integrating Naturalized Metaphysics and Voluntarist Epistemology*. Oxford: Oxford University Press.

Chang, H. (2001). How to Take Realism Beyond Foot-Stamping. *Philosophy*, *76*(1), 5–30.

Chang, H. (2003). Preservative realism and its discontents: Revisiting caloric. *Philosophy of Science*, 70(5):902–912.

Chang, H. (2009). Ontological Principles and the Intelligibility of Epistemic Activities. In *Scientific Understanding: Philosophical Perspectives* (pp. 64–82). University of Pittsburgh Press. http://www.jstor.org/stable/j.ctt9qh59s.7

De Regt, H. W. (2017). *Understanding Scientific Understanding*. Oxford University Press.

De Regt, H. W., & Gijsbers, V. (2017). How false theories can yield genuine understanding. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science* (pp. 50–75). New York: Routledge.

Dellsén, F. (2017). Understanding without Justification or Belief. *Ratio*, *30*(3), 239–254.

Dellsén, F. (2019). Rational understanding: Toward a probabilistic epistemology of acceptability. *Synthese*. https://doi.org/10.1007/s11229-019-02224-7

Dellsén, F. (2020). Beyond Explanation: Understanding as Dependency Modelling. *The British Journal for the Philosophy of Science*, *71*(4), 1261–1286. https://doi.org/10.1093/bjps/axy058

Einstein, A., Podolsky, B., & Rosen, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, *47*(10), 777–780.

Elgin, C. Z. (2017). *True Enough*. MIT Press.

Falguera, J., & de Donato Rodríguez, X. (2016). Flogisto versus oxígeno: Una nueva reconstrucción y su fundamentación histórica. *Crítica (México D. F. En Línea)*, *48*, 87–116. https://doi.org/10.22201/iifs.18704905e.2016.237

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, *71*, 742–752.

Glennan, S. (2017). *The New Mechanical Philosophy*. Oxford: Oxford University Press.

Glennan, S., & Illari, P. (Eds.). (2018). *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. London: Routledge.

Goldstein, H., Poole, C., & Safko, J. (2001). *Classical Mechanics* (3rd ed.). Addison Wesley.

Greco, J. (2014). *Episteme: Knowledge and Understanding* (pp. 284–301). https://doi.org/10.1093/acprof:oso/9780199645541.003.0014

Grimm, S. R. (2008). Explanatory Inquiry and the Need for Explanation. *The British Journal for the Philosophy of Science*, *59*(3), 481–497.

Grimm, S. R. (2011). Understanding. In D. P. S. Berneker (Ed.), *The Routledge Companion to Epistemology*. Routledge.

Grimm, S. R. (2021). Understanding. *Stanford Encyclopedia of Philosophy*.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175.

Henry, J. (2017, February 6). *Newton and Action at a Distance*. The Oxford Handbook of Newton. https://doi.org/10.1093/oxfordhb/9780199930418.013.17

Hills, A. (2016). Understanding Why. *Noûs*, *50*(4), 661–688. https://doi.org/10.1111/nous.12092

Hitchcock, C., & Sober, E. (2004). Prediction Versus Accommodation and the Risk of Overfitting. *The British Journal for the Philosophy of Science*, *55*(1), 1–34. https://doi.org/10.1093/bjps/55.1.1

Hubert, M. (2021). Understanding Physics: `What?', `Why?', and `How?'. *European Journal for Philosophy of Science*, *Forthcoming*.

Kelp, C. (2015). Understanding phenomena. *Synthese*, *192*(12), 3799–3816. https://doi.org/10.1007/s11229-014-0616-x

Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge*. Cambridge University Press. https://doi.org/10.1017/9781108164276

Ladyman, J. (2011). Structural realism versus standard scientific realism: The case of phlogiston and dephlogisticated air. *Synthese*, *180*(2), 87–101. https://doi.org/10.1007/s11229-009-9607-8

Lange, M. (2002). *An Introduction to the Philosophy of Physics: Locality, Fields, Energy, and Mass*. Oxford: Blackwell.

Lawler, I. (2019). Levels of reasons why and answers to why questions. *Philosophy of Science*, 86(1):168–177.

Le Bihan, S. (2017). Enlightening Falsehoods: A Modal View of Scientific Understanding. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining Understanding* (pp. 111–135). New York: Routledge.

Le Bihan, S. (2019). Partial truth versus felicitous falsehoods. *Synthese*, 198:5415–5436.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1–25.

Malfatti, F. I. (2021). On Understanding and Testimony. *Erkenntnis,* 86:1345–1365.

Mizrahi, M. (2012). Idealizations and scientific understanding. *Philosophical Studies*, 160(2):237–252.

Maudlin, T. (2014). What Bell Did. *Journal of Physics A: Mathematical and Theoretical*, *47*(42).

Mill, J. S. (1843). *A System of Logic*. London: George Routledge and Sons.

Mill, J. S. (1859). On Liberty. In M. Philip & F. Rosen (Eds.), *On Liberty, Utilitariansm and Other Essays* (pp. 5–112). Oxford: Oxford University Press.

Morgan, M., & Morrison, M. (Eds.). (1999). *Models as mediators: Perspectives on Natural and Social Science*. Cambridge, UK: Cambridge University Press.

Nawar, T. (2019). Veritism refuted? Understanding, idealization, and the facts. *Synthese,* 198: 4295–4313.

Newton, I. (1999). *The Principia: Mathematical Principles of Natural Philosophy*. Berkeley: University of California Press.

Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago: The University of Chicago Press.

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Regt, H. de, & Baumberger, C. (2020). What Is Scientific Understanding and How Can It Be Achieved? In K. McCain & K. Kampourakis (Eds.), *What Is Scientific Knowledge? An Introduction to Contemporary Epistemology of Science*. New York: Routledge.

Regt, H. W. de. (2009). The Epistemic Value of Understanding. *Philosophy of Science*, *76*(5), 585–597.

Regt, H. W. de, & Höhl, A. E. (2020). Book Review: Understanding, Explanation, and Scientific Knowledge by Kareem Khalifa. *The British Journal for the Philosophy of Science*. https://www.thebsps.org/reviewofbooks/kareem-khalifa-understanding-explanation-and-scientific-knowledge-reviewed-by-de-regt-hohl/

Rice, C. (2019). Understanding realism. *Synthese,* 198:4097–412.

Salmon, W. C. (1998). The Importance of Scientific Understanding. In *Causality and Explanation* (pp. 79–91). Oxford: Oxford University Press.

Schurz, G. (2009). When Empirical Success Implies Theoretical Reference: A Structural Correspondence Theorem. *The British Journal for the Philosophy of Science*, *60*, 101–133.

Schurz, G., & Lambert, K. (1994). Outline of a Theory of Scientific Understanding. *Synthese*, *101*(1), 65–120.

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Sullivan, E., & Khalifa, K. (2019). Idealizations and Understanding: Much Ado About Nothing? *Australasian Journal of Philosophy*, *97*(4), 673–689.

Whewell, W. (1840). *History of the Inductive Sciences*. London: John W. Parker and Son.

Wilkenfeld, D. (2013). Understanding as representation manipulability. *Synthese*, *190*, 997–1016.

Wilkenfeld, D. A. (2017). MUDdy understanding. *Synthese*, *194*(4), 1273–1293.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Woodward, J. F. (2011). Data and phenomena: A restatement and defense. *Synthese*, *182*(1), 165–179.

Ylikoski, P. (2019). Review of "Understanding, Explanation, and Scientific Knowledge" by Kareem Khalifa. *Notre Dame Philosophical Reviews*.

Zagzebski, L. (2001). Recovering Understanding. In M. Steup (Ed.), *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue*. New York: Oxford University Press.