BALKAN JOURNAL
OF PHILOSOPHY

Vol. 9, Issue 1, 2017

The issue is sponsored by the Bulgarian National Science Fund, competition “Bulgarian Scientific Periodicals – 2016”, project No. DNP 05/31 from 21 Dec. 2016

CONTENTS

Special issue devoted to the topic of “Explanation and understanding across the sciences, humanities, and arts”

Interview

Book review

Information for contributors

SPECIAL ISSUE OF "BALKAN JOURNAL OF PHILOSOPHY"

EXPLANATION AND UNDERSTANDING ACROSS THE SCIENCES, HUMANITIES, AND ARTS

The distinction between explanation and understanding was first thematized by Wilhelm Dilthey (1894), who used it to explicate and justify the methodological difference between natural sciences (Naturwissenschaften) and human sciences (Geisteswissenschaften).

Since then, many influential thinkers of the 20th century have referred to this distinction in order to either redefine it and put it to a different use (the neo-Kantians, Weber, Heidegger, Gadamer, Ricoeur, Wittgenstein, von Wright), or to denounce it as misleading, if not completely false (the neo-positivists, Popper, Hempel, Rorty). The 20th century thinkers left us with a great variety of reflections on the topic, and it is to be stressed that the line of demarcation between the two sides of the explanation-understanding divide cannot be neatly correlated with the recognized boundary between analytic and continental philosophy.

The first decade of the 21st century witnessed some notable contributions to this debate. Some cognitive psychologists (Keil) claimed to have provided evidence for the psychological reality of the distinction between inferential explanation and non-inferential direct understanding. At the same time, a new generation of philosophers of science (Faye, Khalifa) distinguished themselves from both of the older camps by taking seriously the distinction between explanation and understanding and attempting to explicate the connection between them in a sensible way. Meanwhile, discussions about the methodological merits of the explanation-understanding distinction continued in areas as different as history, economics, psychiatry, and literature. Such a broad usage of a conceptual distinction across different disciplines and intellectual traditions seems to be a sign that it stands for something truly important. The aim of this Special Issue of the Balkan Journal of Philosophy is therefore to throw light on the significance of the explanation-understanding conceptual distinction in different areas of the sciences, humanities, and arts.

From the Editorial Board of
Balkan Journal of Philosophy

Articles

Invited paper

WORLD-INVOLVING SCIENTIFIC UNDERSTANDING

Stathis Psillos
Dept of Philosophy and History of Science
University of Athens
University Campus
15771 Athens, Greece
e-mail: psillos@phs.uoa.gr

Abstract:
Some philosophers argue that tying scientific understanding to explanation and truth generates a dilemma for a realist view of science: given the practice and the history of science, either we should give up the idea that understanding requires truth, or we should accept that we do not have scientific understanding. Given, we were told, that the latter horn is repugnant, we should jettison the first horn and disconnect understanding and truth. In this paper, I argue that the alleged dilemma for realism is a non-starter. We can accept both that understanding requires truth and claim that theories offer understanding. I argue that understanding and explanation in science go hand in hand, and that—for a realist at least—they should both be world-involving.

Key words: Understanding, explanation, realism, Newton, Einstein .

But science can only be created by those
who are thoroughly imbued with
the aspiration toward truth and understanding.
Albert Einstein 1939

1. Introduction

For many philosophers of science, there is a close link between explanation and understanding. At least prima facie, to understand the reason why X happened is to explain why it happened. Conversely, to explain why X happened conveys understanding of the reason X is the case. But there is also a close link between explanation and truth. At least prima facie, an explanation of why X happened must be true (or truthlike). Putting understanding, explanation, and truth together, however, seems to create a tension. It seems that understanding might be achieved even by strictly false beliefs. But if explanation requires truth and understanding requires explanation, it follows that understanding requires truth. Hence the tension.

The tension is perceived to be especially pressing for a realist view of science. For, some say, scientific realism implies or requires that science offers understanding of the world. But both the practice and the history of science show that understanding can be achieved by false beliefs (theories, models, and what have you). Hence, we should either go with the idea that understanding does not require truth (and sever the link between understanding and explanation), or we should abandon the idea that science offers understanding of the world. This dilemma for realism has been put in a pressing form by Henk de Regt (2015). As he put it: “we face the dilemma of either giving up the idea that understanding requires truth or allowing for the possibility that in many if not all practical cases we do not have scientific understanding” (2015, 3782). On his view, it is preferable to disconnect understanding and truth and to think of understating as a skill rather than as a kind of knowledge (e.g., true explanatory knowledge). More specifically, he defends what I think is a neo-instrumentalist account of scientific understanding, according to which “understanding phenomena consists in the ability to use a theory to generate predictions of the target system's behavior".

I want to resist this neo-instrumentalist view of scientific understanding. I side with Michael Strevens (2013) and Kareem Khalifa (2012) (among others), who defend a close connection between understanding and explanation. But in this paper, my focus will be on the foregoing dilemma. Do realists face a pressing problem with their claim that science offers understanding of the world, the avoidance of which requires severing the link between understanding, explanation, and truth? I will show that they do not, and that the view that scientific understanding is, in some significant sense, tied to world-tracking explanations can be defended by looking at the history and practice of science.

The road map is as follows. In section 2, I will raise the question: can theories offer understanding of the world without at the same time offering explanations of the relevant worldly phenomena? The answer I will defend in the rest of the paper is negative: no explanation in, no understanding out. In section 3, I will argue that the quantity and quality of understanding is a function of how deep the explanation offered by a theory is. I will also introduce the regulative ideal of the "full account" of a phenomenon X, and discuss the relation between understanding and a less-than-full explanation. In section 4, I will argue using Newton and Huygens that an explanation need not be full to be world-involving. In section 5, I will discuss the different modes of understanding and show that they don't pose a problem for realism. Finally, in section 6, I will try to disarm de Regt's dilemma for realism by looking into the transition from Newton's theory of gravity to Einstein's.

2. Understanding via theories

In thinking about scientific understanding, I take my cue from Wesley Salmon (1998, 90), who stressed that

> Understanding in the scientific sense involves the development of a world-picture, including knowledge of the basic mechanisms according to which it operates, that is based on objective evidence—one that we have good reason to suppose actually represents, more or less accurately, the way the world is.

Currently, this view has been under rather heavy fire by philosophers who see scientific understanding as distinct from scientific explanation, and in ways that go far beyond distinguishing between the act of explaining (which is clearly context- and subject-dependent) and what-does-the-explaining (which is an objective worldly feature). In their edited collection *Scientific Understanding: Philosophical Perspectives*, de Regt, Sabina Leonelli, and Kai Eigner (2009) focus on understanding as a cognitive achievement, which is relational (in that it involves an internal mental state of the cognising subject); context-dependent (in that the cognising subject who comes to possess understanding may possess more or less of it, according to what else it is assumed that s/he knows); and a matter of ability (in that it requires extracting understanding from explanation). Antigone Nounou and I have argued in our review (2012) that this psychological approach to scientific understanding is fully compatible with the standard story that scientific understanding and scientific explanation go hand in hand. It can be taken to highlight the significant fact that possession and communication of information regarding what does the explaining is a process involving a rather complex net of abilities, skills, aptitudes, context-related assumptions, etc., without undermining the traditional view that understanding is tied to explanation.

I won't say more on this in this paper, except that when it comes to scientific realism and understanding, the *key* issue is not what kind of understanding a certain subject (a scientist or any other subject) possesses. This is an issue, but not the *key* one. The subject brings in various cognitive constraints. There can be more or less understanding of X by subject S depending on the cognitive capacities, dispositions, and background knowledge of S. And there can be more understanding of X by subject S than by subject S'. These are all aspects of the psychology of understanding. But they are not quite relevant to the issue of scientific realism. Rather, what is central to realism the kind of understanding (of the world) that *science* offers and what it consists of. To address this issue, we must look into what (at least some) scientists say of the world—but the relevant way to proceed is via the scientific theories they espouse. It is scientific theories that offer scientific understanding of the world.

To see the difference between the subject's point of view and the point of view of theories, let us take a brief look at Laplace's monumental work *Traité de Mécanique Céleste*, in which Laplace aimed to offer a rigorous, systematic, and unified account of Newton's theory of gravity. As he put it, since the "epoch" in which "Newton published his discovery of universal gravitation,… Mathematicians have… succeeded in reducing to this great law of nature all the known phenomena of the system of the world, and have thus given to the theories of the heavenly bodies, and to astronomical tables, an unexpected degree of precision" (1829, 21). Laplace's task was "to present a connected view of these theories, which are now scattered in a great number of works." This would be a unification, as he noted in the Preface of the book, of "the whole of the results of gravitation, upon the equilibrium and motions of the fluid and solid bodies, which compose the solar system, and the similar systems, existing in the immensity of space" (ibid.). Laplace's idea was that by means of a rigorous mathematical presentation of the principles of mechanics—what he called "analysis"—he would achieve not only a solution to the problems of astronomy, but also a *theoretical account* of "the application of the principles of mechanics to the motions and figures of the heavenly bodies."

Now, contrast this approach to understanding qua a unified explanation of a body of phenomena with what was noted in the Introduction to the English translation of the book by Nathaniel Bowditch, an esteemed American mathematician. He felt the need to offer some "explanatory notes" which aimed to render Laplace's theory "more accessible to persons who have been unable to prepare themselves for this study by a previous course of reading, in those modern publications, which contain the many important discoveries in analysis made since the time of Newton" (1829, 3). Bowditch felt this need because "from the abridged manner in which [Laplace's] analytical calculations have been made, it has been found difficult to be understood by many persons, who have a strong and decided taste for mathematical studies, on account of the time and labour required." Hence he set for himself the task of "insert[ing] the intermediate steps of the demonstrations necessary to enable them easily to follow [Laplace] in his reasoning."

Clearly, the understanding of the world that the theory offers (by a unified explanation of all relevant phenomena in Laplace's case) is one thing, but the understanding of *this* theory by various subjects (including other scientists) is quite another thing. Understanding the world by means of theories is about finding out what the world is like according to the theory and why it is the way it is; understanding the theory is a matter of intellectual and cognitive skills and capacities.

When scientific realists talk of *science offering understanding of the world*, I take them to be referring to the first kind of understanding (by means of theories), and not to the second kind (understanding the theory by subject X).[1] Laplace's example, which shows something very typical in science, only aims to illustrate why my focus will be on scientific theories and on the question of what kind of understanding of the world *theories* offer. In particular, can theories offer understanding of the world without at the same time offering explanations of the relevant worldly phenomena? The (realist) answer I will defend is negative: no explanation in, no understanding out.

Incidentally, this is not *just* the realist answer. Take a look at Duhem's final pages of the second edition of his masterpiece *The Aim and Structure of Physical Theory* (1914). Duhem was no realist. And yet in trying to dissociate his nuanced approach to science from positivism, he himself felt the need to stress that understanding is tied to explanation, and in particular in getting to the world, in some sense, right. For Duhem, too, scientific theories are not just means of cataloguing information amassed through experiments. Rather, they aim to promote understanding of the world. As he put it, the physicist

> cannot make up his mind to see in physical theory merely a set of practical procedures and a rack filled with tools…. [H]e cannot believe that it merely classifies information accumulated by empirical science without transforming in any way the nature of these facts or without impressing on them a character which experiment alone would not have engraved on it. If there were in physical theory only what his own criticism made him discover in it, he would stop devoting his time and efforts to a work of such a meagre importance (1906, 334).

His very idea that the classification offered by scientific theories should be *natural* is meant to distance himself from the view that understanding is not connected with getting at least *some* facts of the world right. Scientific understanding, as I shall claim, should be world-involving.

[1] This is akin, but not identical with, what Strevens (2015) has called "understanding with" theories.

3. Levels of understanding are levels of explanation

The existence of different levels (or depths) of understanding is something widely acknowledged. It is customary to say that understanding comes in degrees—there can be more or less of it. But whereas this is typically taken to be a cognitive achievement of a subject, I have a different approach: scientific theories can offer more or less understanding of the way the world is and of how various phenomena are produced and are connected to each other. The quantity and quality of understanding is a function of how deep the explanation offered by a theory is. So scientific understanding is indeed a cognitive achievement, but an achievement, if I may say so, of a *theory*. It is a theory (or a model, or some other representational device for that matter) that offers understanding of the world, and it can offer more or less of it. In fact, this understanding might well be cumulative in the sense that newer theories offer more understanding of their intended domain than their predecessors. But even among contemporaneous theories, the situation can be such that one of them offers more understanding of certain phenomena than its rivals.

Take, for instance, Maxwell's theory of the electromagnetic field. By the time he developed his unified theory, the domain of electricity and magnetism was, in Helmholtz's words a "pathless wilderness" (1899, x). Circa 1840, there were a number of rival theories and competing hypotheses concerning electricity and magnetism. "This plentiful crop of hypotheses," as Helmholtz put it, "had become very unmanageable, and in dealing with them it was necessary to go through complicated calculations, resolutions of forces into their components in various directions, and so on… Observed facts and deductions from exceedingly doubtful theories were inextricably mixed up together." Order was restored when Maxwell introduced the notion of the electromagnetic field and developed (a form of) the equations that capture the fundamental laws of behaviour of the field.

Maxwell's unification of electricity and magnetism and the subsumption of optical phenomena under it offered more understanding of optical phenomena than the theories of optics before Maxwell. The reason is simple: it *explained* them (in a more truthlike way). More specifically, whereas before Maxwell, the optical phenomena were taken to be *sui generis* (and hence, the demand was that they be understood in their own terms), after Maxwell's theory they were taken to be instances of a wider class of electromagnetic phenomena which had to be understood in their own terms. Hence, understanding optical phenomena did not require a separate theory. Instead, the theory that—if it were true—would explain electromagnetic phenomena, would also explain optical phenomena. This is not a matter of less cognitive effort; it is a matter of conceptual and explanatory unification. The same point can be said when it comes to the most recent development, viz., the unification of electromagnetic and weak interactions. Ultimately, the improvement in understanding has to do with the fact that understanding requires some (at least temporarily and perhaps ultimately) unexplained explainers. If explanation were to go back *ad infinitum,* as it were, there would be no understanding, because whether X is understood would be pushed back to whether Y is understood, and so on without end. The fewer the unexplained explainers, the better the conceptual and explanatory economy achieved; hence the better the theoretical understanding.[2]

If there can be more or less (or a better and worse) understanding of X, there should be (at least ideally and in principle) a "full understanding" of X. I take it that see the full understanding of X as the full "account" of X. We could say that the full "account" of X is the full (complete) *explanation* of X—of why X is the case. This might never be achieved, but the realist account of science presupposes that for each and every natural phenomenon X, there is a full account of it (and hence an explanation of it). Note that this is meant to be an external constraint on understanding. It does not presuppose that there will be a single set of conceptual means to grasp it or map it out. Nor does it presuppose a certain view of causation, or of explanation. It might be, for instance, that there are genuinely stochastic happenings in the world. The full account of them will then be whatever in the world makes them happen (even if this is an irreducibly stochastic process). Hence, by "full account" of natural phenomenon X, I mean whatever it is in the world that makes X be the case. If nothing in the world is such that it makes

[2] Though the modern locus classicus of this view is Friedman's "Explanation and Scientific Understanding" (1974), the basic idea can be found in Feigl (1970, 12). He takes "the fact-postulate ratio" to represent "the explanatory power of theories." As he put it: "The aim of scientific explanation throughout the ages has been *unification*, i.e., the comprehending of a maximum of facts and regularities in terms of a minimum of theoretical concepts and assumptions."

X to be the case (if X is a brute fact, as it were), then so be it—that is the full account, whether we find it or not. If, on the other hand, there is a complicated network of laws, causes and particular matters of fact that make X be the case, so be it—the "full account" of X would be (the mapping of) this network. To possess the full account of X is to possess the complete explanation of X. In this sense, to understand X *fully* is for science to possess (and to offer) the complete explanation of X.

All this is akin to what Peter Railton has called "ideal explanatory text." Railton proposed that the occurrence of a certain event, be it likely or not, is explained by placing this event within a web of

> inter-connected series of law-based accounts of all the nodes and links in the causal network culminating in the explanandum, complete with a fully detailed description of the causal mechanisms involved and theoretical derivations of all covering laws involved (1981, 174).

As is well-known, he tied explanation to a Deductive-Nomological-Probabilistic (DNP) text. My main divergence from Railton's view is that I take the full explanatory account of a phenomenon X to be the worldly laws and processes that make X be the case, rather than the ways this explanatory account is represented. Hence, though Railton may well be right in representing this full account as a DNP text, this is an issue that may be left as an open question; whether or not the full explanatory account may be mapped by a DNP structure is of secondary importance. Of primary importance is the *ontic* issue of the very existence of the full account of X.

Getting the full explanatory account of phenomenon X is what scientific understanding by means of theories amounts to—at least for realists. But it is a truism that, for all practical and theoretical purposes, this is an idealization (or a regulative ideal, if you prefer). As Railton himself pointed out, in practice what scientists need to (and should) strive for is "explanatory information" relevant to the explanandum. Such information, if indeed it is information *relevant* to the explanandum, should be part of the full account (or the ideal DNP text in Railton's case). But it should be clear that scientific understanding, in the sense of fullness noted above, is not achievable—perhaps not even in the ideal limit of inquiry. The reason is familiar, and has to do with the immense complexity of the world and the inevitability of abstraction and idealization in scientific representations of worldly phenomena. To explain X requires representing both X and the explanans of X, and since representation inevitably involves idealization and abstraction, explanation itself (i.e., acquiring explanatory information) relies on idealization and abstraction.

Let's agree that it is rather unlikely (if possible at all) that science offers us "full understanding" of the world. This does not mean that it cannot offer us a partial but deep (and increasingly more accurate) understanding of the world. The understanding of the world offered by science can be genuine even if it is not full. *Understanding by omission* is still understanding, if the non-omitted explanatory information is the one that contributes to (or is required for) the explanation of the relevant aspects of X. For instance, for understanding the orbit of Mars—that is, possessing genuine explanatorily relevant information for the shape of its orbit—there is no need to understand or know that Mars is red. *Understanding by distortion* is still understanding, if the distorted explanatory information (e.g., that Mars is a mass-point) is such that it captures the causally relevant aspects of X; that is, the information that is required for accounting for the basic features of the phenomenon under study. Some omissions and distortions can be put back in, as it were, since there is still room for a better understanding of the orbit of Mars by making the explanation more accurate and more in tune with the actual orbit of Mars. Hence, whereas initially the perturbations of Mars' orbit by other planets might well be omitted (without affecting the understanding of Mars' orbit) a "fuller" understanding is achieved by incorporating into the explanation of Mars' orbit second-order phenomena (more on this in a while).

Newton is a case in point. Take Proposition 1, Theorem 1 of Book 1 of the *Principia*: "The areas which revolving bodies describe by radii drawn to an immovable centre of force do lie in the same immovable planes, and are proportional to the times in which they are described" (1729, 40). Newton offered a purely geometrical proof applied to objects (*revolving bodies*) whose only properties are those relevant to their motion (revolution). The law (or half of the law, to be precise: if the area-law is satisfied, then there is a centripetal force) expresses an exact relation among magnitudes: revolving bodies, radii, immovable centre, immovable plane, force. But Newton was fully aware that "the sun gravitates towards all the planets, and all the planets towards the sun. For the secondary planets, while they ac-company the primary, revolve the meanwhile with the primary about the sun. Therefore, by the same argument, the planets of both kinds gravitate towards the sun and the sun towards them" (Roger Cotes, Preface to the second edition, 1729, xxv). Hence, while Proposition 1 is about "an immovable Centre of force," it is the case that "the sun gravitates towards all the planets"; hence it is not immovable.

Was this something that gave Newton pause? To cut a long story short, it was certainly *not*. The laws apply to the phenomena "*quam proxime*." This phrase (which occurs 139 times in the *Principia*) literally means "most nearly to the highest degree possible"; it is probably best translated "very, very nearly," or, as is more customary, "very nearly." Hence, though Proposition 1, Theorem 1 of Book 1 does not apply directly and exactly to the relevant phenomena, it does explain them, and it does offer understanding of them, since the actual concrete physical system is like the one described by the law-statement, *quam proxime*.

As George Smith noted, Newtonian idealizations (that is, Newton's theory itself) make possible the following: on the assumption that the theory is exact (that is, on the basis of idealizations embodied in the theory), there are systematic discrepancies between the predictions of the theory and the observations. These discrepancies constitute "second order phenomena" which the theory should explain by attributing them, as far as possible, to physical causes. Hence, the theory is developed in such a way that a "deeper understanding" of worldly phenomena is achieved by finding further *explanations* of the discrepancies between predicted values and observed values of various parameters. The theory is being further elaborated and tested by finding the causes of discrepancies. This is achieved by "asking in a sequence of successive approximations, what further forces or density variations are affecting the actual situation?" (2014, 277).

One reason why Smith has called this kind of phenomena "second order" is that their being acknowledged as phenomena worthy of accommodation *presupposes* the (at least partial) explanatory correctness of the theory that makes them available. In the case of planetary motion, the orbital discrepancies which call for a physical explanation presuppose Newton's theory of gravity. They are made available qua phenomena as discrepancies between the idealized predictions of Newton's theory (which hold on the assumption that it is an exact theory) and observations. As Smith put it: "they are what you get by subtracting observations from idealized calculated results." And he added: "They are second-order because they categorically presuppose the theory of gravity, taken as holding exactly. They are phenomena because they are systematic and hence constitute regularities that cannot initially be identified because more dominant regularities mask them" (2014, 278).

Not only is understanding of the planetary motions achieved by Newton's theory, since at least some of its principles hold *quam proxime* in their intended domain; but also the theory itself guided further and deeper understanding of the relevant phenomena by developing theoretical means with which deviations from the predictions of the theory are to be *explained*.

4. The limits of fullness

Understanding requires some of the "full account" of X. And unless otherwise informed, to possess more of the full account of X is to possess more information about why X is the case. This is not meant to imply that a "deeper understanding" of X is always necessary. But it does imply that where to stop in looking for understanding is a function of where to stop looking for explanation. Some understanding of the reason why X is the case might well already be deep enough without the need to search for more of the "full account."

Newton, again, is a case in point. In his preface to the first edition of the *Principia*, he summarized what he was doing in the third book thus (1729, xxxiv):

> In the third Book I give an example of this in the *explication* of the System of the World; for by the propositions mathematically demonstrated in the former Books, in the third I derive from the celestial phenomena the forces of gravity with which bodies tend to the sun and the several planets. Then from these forces, by other propositions which are also mathematical, I deduce the motions of the planets, the comets, the moon, and the sea (emphasis added).

Not only did Newton aim to explain the System of the World; it is also fair to say that it was by means of this explanation that he achieved a deep understanding of gravity (certainly deeper than his contemporaries, as we are about to see). And yet as he admitted, he did not assign a cause to gravity. In his famous words:

> Hitherto we have explained the phenomena of the heavens and of our sea by the power of gravity, but have not yet assigned the cause of this power. This is certain, that it must proceed from a cause that penetrates to the very centres of the sun and planets, without suffering the least diminution of its force; that operates not according to the quantity of the surfaces of the particles upon which it acts (as mechanical causes used to do), but according to the quantity of the solid matter which they contain, and propagates its virtue on all sides to immense distances, decreasing always as the inverse square of the distances (1729, vol.2, 546).

Would assigning a cause to gravity offer a *deeper* explanation of gravity? It seems that Newton himself thought so. It also seems that most (if not all) of his contemporaries thought so, too. Take Huygens, for instance. In his *Discourse on the Cause of Gravity* (1690) he expressed his dissatisfaction with Newton's failure to offer a *mechanical* explanation of the cause of gravitational attraction. But he had no difficulty in granting that Newton's law of gravity was essentially correct when it comes to accounting for the planetary system. As he put it:

> I have nothing against *Vis Centripeta*, as Mr. Newton calls it, which causes the planets to weigh (or gravitate) toward the Sun, and the Moon toward the Earth, but here I remain in agreement without difficulty because not only do we know through experience that there is such a manner of attraction or impulse in nature, but also that it is explained by the laws of motion, as we have seen in what I wrote above on gravity (1997, 31).

Explaining the fact that gravity depends on the masses and diminishes with distance "in inverse proportion to the squares of the distances from the centre" were, for Huygens, clear achievements of Newton's theory, despite the fact that the mechanical cause of gravity remained unidentified.

For Huygens, the full understanding of a natural phenomenon was tied to a certain *mode of understanding*, viz., mechanical explanation. He said, regarding Descartes:

> "Mr Descartes has recognized, better than those that preceded him, that nothing will be ever understood in physics except what can be made to depend on principles that do not exceed the reach of out spirit, such as those that depend on bodies, deprived of qualities, and their motions" (1997, 1–2).

This is a rather substantial issue. For one, even if one were to agree that this is the proper mode of understanding of natural phenomena, one need not necessarily agree that a less-than-full explanation is a bad or inadequate explanation. We saw Huygens himself did not think this; nor did Newton. For another, one may disagree that there is only one proper mode of understanding in science. In the case of Huygens and Newton, what was at stake, at least partly, was the issue of the limits of explanation. Newton thought that explanation need not be full to be adequate: looking for a fuller explanation is not always the right attitude if the one already offered is satisfactory. For instance, a fuller explanation of gravity would have to be, Newton thought, hypothetical and disconnected from what is licensed by his own experimental method. In his famous words:

> But hitherto I have not been able to discover the cause of those properties of gravity from phenomena, and I frame no hypotheses; for whatever is not deduced from the phenomena is to be called an hypothesis; and hypotheses, whether metaphysical or physical, whether of occult qualities or mechanical, have no place in experimental philosophy. In this philosophy particular propositions are inferred from the phenomena, and afterwards rendered general by induction. Thus it was that the impenetrability, the mobility, and the impulsive force of bodies, and the laws of motion and of gravitation, were discovered. And to us it is enough that gravity does really exist, and act according to the laws which we have explained, and abundantly serves to account for all the motions of the celestial bodies, and of our sea (1729, vol.2, 547).

Note that the issue here is *not* understanding vs explanation. Rather, the issue is that the understanding offered by a theory need not be the full explanation of the relevant phenomena. What counts as enough explanation for understanding is an open issue, which has to do with substantive questions concerning scientific theories and their scope and limits. Explaining why a certain phenomenon holds is certainly part of the understanding of the phenomenon offered by the relevant theory. But how deep this understanding should be depends a lot on substantive issues concerning the *modes of understanding*.

5. Modes of understanding

By "modes of understanding," I mean the specific forms which scientists think that explanation, and hence understanding, must take in order to be considered genuine explanations which offer genuine understanding. The mode of understanding is typically under substantive dispute among advocates of different theories. Like Huygens, Leibniz, too, objected to Newton's theory of gravity on the grounds that it does not offer a mechanical explanation of gravity. He was very clear that though he allowed "magnetic, elastic, and other sorts of forces," they are permissible "only insofar as we understand that they are not primitive or incapable of being explained, but arise from motions and shapes" (1989, 313). So for Leibniz forces are necessary for explanation, but a condition for their applicability to the natural world is that they are not primitive, but "arising from motions and shapes." Leibniz's worry was that forces acting at a distance "lead us back to the kingdom of darkness," viz., occult qualities. This is "barbarism in physics," as he put it: to posit essential forces to matter which are superadded to it and ground "the general striving [nisus] of matter for matter" (1989, 314). Newtonian gravity was taken to such a barbaric force! For Leibniz the preferred mode of explanation—and hence of understanding—of natural phenomena was via positing an intelligible mechanism which underscores the natural forces.

In a letter to Hartsoeker (Hanover, 10 February 1711), Leibniz made clear that without a mechanism, this force is "an unreasonable occult quality." He said:

> Thus the ancients and the moderns, who own that gravity is an occult quality, are in the right, if they mean by it that there is a certain mechanism unknown to them, whereby all bodies tend towards the center of the earth. But if they mean that the thing is performed without any mechanism by a simple primitive quality, or by a law of God, who produces that effect without using any intelligible means, it is an unreasonable occult quality, and so very occult, that it is impossible it should ever be clear, though an angel, or God himself, should undertake to explain it (in Newton 2004, 112).

Newton, on the other hand, insisted on a mode of understanding according to which it is enough to state the law by means of which gravity operates. In an unsent letter written circa May 1712 to the editor of the Memoirs of Literature, Newton referred explicitly to Leibniz's letter to Hartsoeker, and stressed that it is not necessary for the introduction of a power—such as gravity—to specify anything other than the law it obeys:

> And therefore if any man should say that bodies attract one another by a power whose cause is unknown to us, or by a power seated in the frame of nature by the will of God, or by a power seated in a substance in which bodies move and float without resistance and which has therefore no *vis inertiae* but acts by other laws than those that are mechanical: I know not why he should be said to introduce miracles and occult qualities and fictions into the world (2004, 116).

Note well Newton's point. The fact that an explanation does not conform to a certain mode of understanding does not make it fictitious, occult, or miraculous. There are simply multiple modes of understanding.

Towards the end of the nineteenth century, there was an analogous debate concerning electromagnetic phenomena. Famously, William Thomson (Lord Kelvin) noted: "I never satisfy myself until I can make a mechanical model of a thing. If I can make a mechanical model I can understand it. As long as I cannot make a mechanical model all the way through I cannot understand; and that is why I cannot get the electro-magnetic theory" (1987, Lecture XX, 109). He believed that the preferred mode of scientific understanding was via building mechanical models of natural phenomena. But Maxwell was more cautious. He regarded mechanical models as illustrative rather than explanatory. He invited his readers to understand "literally" his claims about the energy of the field (1864, 70), and yet—regarding the mechanical models he had offered—he told his readers *not* to take them literally. He stated that he

> wish(es) merely to direct the mind of the reader to mechanical phenomena which will assist him in understanding the electrical ones. [That is] all these phrases [i.e., related to mechanical representations]... are to be considered as illustrative, not as explanatory (ibid.).

The mature dynamical theory of electromagnetic fields that he developed rested on the general principles of dynamics, and was independent of any particular model concerning electromagnetic fields (cf. Maxwell 1873, Vol.2, chapters 5–9).

This difference in modes of understanding was picked up by Duhem, who—building on Rankine's work—distinguished between "hypothetical" theories and "abstractive" ones (cf. Duhem 1906, 52–3). The former is, roughly, a theory that posits specific explanatory-causal mechanisms of the phenomena, whereas the latter is, roughly, a theory that consists of general principles from which all phenomena are supposed to follow. Interestingly, Rankine himself distinguished between two kinds of hypothetical theories "according as [the conjectural conceptions] are adopted as a probable representation of a state of things which may really exist, though imperceptible to the senses, or merely as a convenient means of expressing the laws of phenomena." He called the first kind "objective" and the second "subjective." Equally interesting is that Rankine did not devalue the hypothetical mode of understanding. He took it to be an indispensable "preliminary step" for the reduction of "the expression of the phenomena to simplicity and order, before it is possible to make any progress in framing an abstractive theory" (1855, 213). For him, the contrast was between two different modes of unification, of the "tendency… to combine all branches of physics into one system." One way is to rely on the axioms of mechanics as "the first principles of the laws of all phenomena; an object for the attainment of which an earnest wish was expressed by Newton," the other being to rely on "propositions comprehending as particular cases the laws of the particular classes of phenomena comprehended under the more extensive classes."

The difference between these two modes of understanding offered by scientific theories is not that one is explanatory while the other is not. Both, being unificatory, are explanatory. The difference, it was thought, was epistemic. Hypothetical understanding is more risky (and hence less certain) than abstractive understanding, since abstractive theories end up with general principles which are "applicable to physical phenomena in general, and which, being framed by induction from facts alone, will be free from the uncertainty

which must always attach, even to those mechanical hypotheses whose consequences are most fully confirmed by experiment" (1855, 213).

This distinction may well resonate with Albert Einstein's distinction between theories of principles and constructive theories. In a piece Einstein wrote for *The Times* of London published on 28 November, 1919 ("What is the Theory of Relativity?"), he drew this distinction as follows:

Constructive theories are "attempt(s) to build up a picture of the more complex phenomena out of the materials of a relatively simple formal scheme from which they start out." A typical example was the atomic theory of matter, which "seeks to reduce mechanical, thermal, and diffusional processes to movements of molecules—i.e., to build them up out of the hypothesis of molecular motion" (1954, 228).

Of the "principle-theories," Einstein said: "The elements which form their basis and starting-point are not hypothetically constructed but empirically discovered ones, general characteristics of natural processes, principles that give rise to mathematically formulated criteria which the separate processes or the theoretical representations of them have to satisfy" (1954, 228). A typical example was his own Special Theory of Relativity.

Constructive Theories rely on hypotheses about microscopic entities and their properties. Principle-Theories avoid hypotheses and start with general principles which have an empirical basis—they can be thought of as laws that cover all relevant phenomena and are grounded in experience.

Einstein was sensitive to the difference between these two types of theories, but endorsed and developed both of them. *Prima facie*, there is an epistemological difference between them, which Einstein put thus: "constructive speculation" increases "the distance between the foundation of the theory and sense experiences" (1954, 305). So Principle Theories seem to stay closer to experience, and hence be more certain than constructive theories. But Einstein was clear that this is not a reason for scepticism. Speculative theories, no less than principle-theories, can achieve "logical unity" and can unify various seemingly disparate phenomena. The prime example again was the atomic theory of matter, which "produced… definite values for the true magnitudes of atoms and molecules which resulted from several independent methods and were thus placed beyond the realm of reasonable doubt" (1954, 303).

Both kinds of theory are explanatory and offer understanding of the world. We may call the principles of constructive theories *speculative* and the principles of principle-theories *comprehensive*. They differ in their scope. Yet both are needed for science; both promote unity; both are testable. Though comprehensive principles are, as a rule, epistemologically more certain than speculative principles (since they stay closer to experience), speculative principles can acquire certainty by their empirical success and their unifying power.

What is important for our purposes is that when it comes to realism, there is no reason to prefer one over the other mode understanding *per se*. The issue is *not* the mode of understanding offered by a scientific theory, but rather whether there are evidential and theoretical reasons to take a certain theory as true or as partially true. In fact, insofar as the accounts offered by the two modes of understanding are not inconsistent with each other, there is no reason to prefer one to the other, provided there is no contrary evidence. It might well be the case that certain instrumentalists put a premium on developing abstract theories. This may well be because abstract theories may be more amenable to an instrumentalist construal as schemes for prediction and control. But this need not be so. There is nothing inherent in abstract theories that makes them instrumentalist, as Einstein's principle-theories amply show. Abstract theories can be explanatory qua unifying and likely to be true qua supported by empirical evidence.

Salmon noted that

> there are at least two intellectual benefits that scientific explanations can confer upon us, namely, (1) a unified world picture and insight into how various phenomena fit into that overall scheme, and (2) knowledge of how things in the world work, that is, of the mechanisms, often hidden, that produce the phenomena we want to understand (1998, 89).

He added that these two kinds of explanation are compatible with each other, and that they can "complement rather than conflict with each other." I think that this is essentially correct. But I would not want to add that it is always necessary to look for both kinds of explanation. There is no ultimate criterion for preferring one mode of understanding to the other, and in the end, the judges should be scientists themselves. Actually, insofar as both modes of understanding capture and offer some information about the full account of the phenomena to be explained, they are both world-involving, even if they are not full.

6. Understanding and truth

There is a crucial question that has been lurking in the shadows: what is the connection between explanation and truth? The bulk of the paper so far has been an attempt to show that explanation based on theories can be genuine and adequate even if it is not strictly true. It is enough for it to be world-involving, that is to be part of the "full account" for phenomenon X. In fact, even if an explanation captures the right causal factors for X, it is unlikely (if even possible) that it is strictly true. This, we noted, is no surprise, given that theories involve idealizations and abstractions. Theoretical explanation requires and relies upon representation, and this is (almost invariably and always) inexact and partial. Understanding is tied to explanation; explanation should be world-involving (even if it is not world-exhausting); hence, understanding should be world-involving. This should be enough for scientific realism.

But what about the argument from the history of science? The falsity of past theoretical explanations of phenomena has been taken to be an argument against the view that understanding is tied to explanation. The claim is that past false theories offered understanding of the world *even if* the explanations offered by them were false (see de Regt 2015, 3789-90). In this final section, I want to reverse this claim. I want to argue that past theories offered understanding of the world because and insofar as the explanations offered were not characteristically false, but world-involving (and hence, in some sense, partially true).

This is not the place to go into the details of the argument from pessimistic induction; suffice it to say that if the realist responses to the argument are on the tight track—and I have argued elsewhere that they are (cf. my 1999; 2009)—then there is a lot more explanatory continuity in theory-change than is admitted by the pessimists. For now, let us only focus on an example that strikes de Regt as indicative of the alleged problems for realism: Newtonian mechanics. He says:

> But the thesis that Newton's theory of gravitation—which is an essential part of current high-school and university physics education—does not give us genuine understanding of phenomena such as tidal motion, planetary behavior, etcetera, because it has been proven false by Einstein and should be replaced by the theory of general relativity, will surely strike many as absurd. Yet the realist thesis does not discriminate between Newtonian gravitational force and phlogiston: neither entity is included in the ontology of contempo-

> rary science. The crucial point is that both developments involved a radical, discontinuous shift from one ontology to another. This implies that the realist thesis applies in neither case: we cannot retain the idea that Newtonian theory is approximately true, while claiming that phlogiston theory is radically false (2015, 3789–90).

Unlike what de Regt asserts, the crucial point is precisely that we *can* retain the idea that the Newtonian theory is world-involving (and approximately true), while claiming that phlogiston theory is radically false. Hence, we can explain why Newton's theory still offers (some) understanding of the world, while the phlogiston theory does not. No one cares about phlogiston anymore simply because, to put it bluntly, phlogiston does not feature in any (let alone the full) account of the phenomena of combustion; it offers no explanatory information whatsoever. Phlogiston-based explanations are not world-involving. Things could not be more different vis-à-vis Newton's *law of gravity*. Newton's account features prominently in the full(er) account of gravitational phenomena because it correctly identified masses as causally relevant factors in the stability of the solar system, and accounted for gravity on this basis. This is something that can be evinced by looking, among other things, at how Einstein himself reacted to Newton's theory.

Einstein never doubted that he had to recover Newton's theory of gravity as a *limiting case* of his own theory. In his letter to Hilbert (18/11/1915), he emphasized the role of the Newtonian limit in his formulation of the theory of relativity. Stressing that he arrived at the field equations first, he noted that three years earlier

> … it was hard to recognize that these equations form a generalisation, and indeed a simple and natural generalisation, of Newton's law. It has just been in the last few weeks that I succeeded in this (I sent you my first communication), whereas 3 years ago with my friend Grossmann I had already taken into consideration the only possible generally covariant equations, which have now been shown to be the correct ones. We had only heavy-heartedly distanced ourselves from it, because it seemed to me that the physical discussion had shown their incompatibility with Newton's law (quoted in Renn et al. 2007, 908)

In his 1916 systematic presentation of GTR, Einstein (1916, 145) made clear that starting from the field equations of gravitation "give us, in combination with the equations of motion… to a first approximation Newton's law of attraction, and to a second approximation the explanation of the motion of the perihelion of the planet Mercury... These facts must, in my opinion, be taken as convincing proof of the correctness of the theory."

If Einstein thought that recovering Newton's law of gravity within his GTR was "convincing proof" of his *own* theory, it should surely be the case that Einstein thought that Newton's theory was part of the full(er) account of gravity he had just offered; i.e., that it is world-involving.

As is well-known, Einstein's chief point was that in a quasi-static and weak gravitational field, he could recover Newton's law of attraction in the form of the gravitational potential. In a quasi-static field, only component g_{44} of the metric tensor g was needed to determine the motion of a material point, and hence to arrive at Newton's law of gravity. Hence the values of $g_{11}…g_{33}$ do not affect the first-order equation of motion, and hence the Newtonian limit. As he put it in a letter to Michele Besso (10/12/1915):

> Most gratifying is the agreement with perihelion motion and the general covariance; strangest, however, is the circumstance that Newton's theory of the field is incorrect already in the 1st order eq. (appearance of the g11…g33). It is just the circumstance that the g11…g33 do not appear in first-order approximations of the motion eqs. which determines the simplicity of Newton's theory (quoted by Renn et al. 2007, 283).

On November 18, 1915, Einstein published a paper in which, as he put it, he found "an important confirmation of this most fundamental theory of relativity, showing that it explains qualitatively and quantitatively the secular rotation of the orbit of Mercury (in the sense of the orbital motion itself), which was discovered by Leverrier and which amounts to 45 sec of arc per century." He was, of course, referring to the General Theory of Relativity, which he had presented a week earlier in November 1915 in the same journal.

In explaining the anomalous perihelion of Mercury, Einstein took it for granted that Newton's theory (assisted by standard perturbation theory) successfully accounted for 531 arc-second per century of the total 575.1 arc-seconds per century precession of the perihelion of Mercury versus the fixed stars. In fact, taking seriously the 43.4 arc-seconds per century as the observed discrepancy that constituted an anomaly for Newton's theory *presupposed* that the remaining 531 arc-second per century gravitational effects from all the other planets on the precession of the perihelion of Mercury were accounted for by Newton's theory. And this required that Newton's theory was solid enough to allow taking *this* theory as the benchmark for the explanation of the 531 arc-second per century.

Starting with the field equations, which express the motion of a material point in a gravitational field in a vacuum,

$$\sum_{\alpha} \frac{\partial \Gamma^{\alpha}_{\mu\nu}}{\partial x_{\alpha}} + \sum_{\alpha\beta} \Gamma^{\alpha}_{\mu\beta}\, \Gamma^{\beta}_{\nu\alpha} = 0$$

Einstein obtained a solution using successive approximations. He then arrived at the equation

$$\left(\frac{dx}{d\varphi}\right)^2 = \frac{2A}{B^2} + \frac{\alpha}{B^2}x - x^2 + \alpha x^3$$

where $x = 1/r$ is the inverse of the radial distance from the Sun, and φ is the angular coordinate in the orbital plane. The symbols A and B are constants of integration. B is related to Kepler's second law ($B = constant = r^2 d\varphi/ds$), and A is related to the energy conservation per unit mass ($A = constant = (1/2)u^2 + \Phi$). Φ is the Newtonian gravitational potential $\Phi = -GMsun/r = -\alpha/2r$. As Einstein noted, this equation "distinguishes itself from the corresponding Newtonian theory only through the last member on the right side." Now, if the factor ax^3 is excluded, the solutions of (1) are Newtonian elliptical orbits. If ax^3 is included, then the solutions are precessing Newtonian elliptical orbits. As he put it: "The calculation yields, for the planet Mercury, a perihelion advance of 43'' per century, while the astronomers assign 45''+–5'' per century as the unexplained difference between observations and the Newtonian theory. This theory therefore agrees completely with the observations." Note that though gravitational forces are not part of this account, the kindred notion of gravitational potential (as formulated by Laplace) plays a key role.

In sum, some key theoretical components of Newton's theory—the law of attraction and the claim that the gravitational effects from the planets on each other were a significant cause of the deviations from their predicted orbits—were taken to be broadly correct and explanatory (of at least part) of the phenomena to be explained. Here is a characteristic way to put the matter:

> Our modern acceptance of general relativity is not only based on experiments or observations related to some of its special predictions but also on the fact that it incorporates the entire Newtonian knowledge on gravitation, including its relation to other physical interactions, that has been accumulated over a long period of time in classical physics and in the special theory of relativity. This knowledge embraces, among other aspects, Newton's law of gravitation including its implications for the conservation of energy and momentum, the relation between gravitation and inertia, the understanding that no physical action can propagate with a speed greater than that of light, which was first achieved by the field theoretic tradition of classical physics and then finally established with the formulation of special relativity, and, more generally, the local properties of space and time, also formulated in special relativity (Renn & Sauer, 303).

The point here is not just that no similar case can be made for phlogiston. I think that de Regt cannot really be serious when he says the following: "For example, why not use eighteenth-century phlogiston theory (which was discarded in favor of Lavoisier's oxygen theory) to understand particular chemical phenomena? My answer is that there is no principled objection to doing so" (2015, 3792). The main point is that understanding requires world-involving explanations, even though the latter might not be full. Insofar as the explanatory information offered by a scientific theory is, in the relevant sense, partially true or truthlike, then the theory offers understanding. From the point of view of the newer theory, this is typically captured by retaining some parts of the preceding theory and jettisoning others. The newer theory will typically offer a deeper understanding of the relevant phenomena because it will offer a deeper and more correct (more world-involving) explanation of them.

7. Final remarks

Recall the dilemma for realism that we started with: given the practice and the history of science, either we should give up the idea that understanding requires truth, or we should accept that we do not have scientific understanding. Given, we were told, that the latter horn is repugnant, we should jettison the first horn and disconnect understanding and truth. But if we do so, I claimed, we forfeit the very idea of understanding as world-involving.

The alleged dilemma for realism is a non-starter. We can accept both that understanding requires truth *and* claim that theories offer understanding. What we have to abandon, if we really need any reminder, is the thought that scientific understanding requires the whole truth and nothing but the truth. Similarly for scientific explanation: it has to involve some parts of the "full account," but even then the proviso should be added that, given that representation is always subject to idealizations and abstractions, explanation will, at best, be *quam proxime*. Understanding and explanation in science go hand in hand, and—for a realist at least—they should both be world-involving.

References

de Regt H W (2015) Scientific Understanding: Truth or Dare? *Synthese* 192: 3781–3797.

de Regt H W, Leonelli S, Eigner K (eds) (2009) *Scientific Understanding: Philosophical Perspectives*. Pittsburgh: University of Pittsburgh Press.

Duhem P (1906) *The Aim and Structure of Physical Theory*. 2nd edition 1914. trans. P. Wiener, Princeton, NJ: Princeton University Press (1954).

Einstein A (1916) The Foundation of the General Theory of Relativity, reprinted in A. Einstein *The Principle of Relativity*. Dover 1932.

Einstein A (1954) Ideas and Opinions. Bonanza Books.

Feigl H (1970) 'Orthodox' View of Theories: Remarks in Defence as Well as Critique. In *Minnesota Studies in the Philosophy of Science*, 4, eds M. Radner, S. Winokur, 3–16. Minneapolis: University of Minnesota Press.

Friedman M (1974) Explanation and Scientific Understanding. *Journal of Philosophy* 71: 5–19.

Helmholtz H (1899) "Preface" in Heinrich Hertz *The Principles of Mechanics: Presented in a New Form*, New York: MacMillan & Co

Huygens Ch (1690) *Discourse on the Cause of Gravity*. Karen Bailey (trans). Mimeographed 1997.

Khalifa K (2012) Inaugurating Understanding or Repackaging Explanation? *Philosophy of Science* 79:15–37.

Laplace, Pierre Simon Marquis de (1829) *Celestial Mechanics*, Volume 1. Translated with a commentary by Nathaniel Bowditch. Boston: Hillard, Gray, Little and Wilkins publishers.

Leibniz G W (1989) Philosophical Essays. Roger Ariew & Daniel Garber (trans.), Hackett Publishing.

Maxwell J C (1864) *A Dynamical Theory of the Electromagnetic Field*. Edited with an Introduction by T. F. Torrance, Edinburgh: Scottish Academic Press, 1982.

Maxwell J C (1873) *A Treatise on Electricity and Magnetism*, Vol. 2, 3rd edn, Oxford: Clarendon Press.

Newton,I (1729) *The Mathematical Principles of Natural Philosophy*. Andrew Motte (trans.) Berkeley and Los Angeles: University of California Press, 1934.

Newton I (2004) *Philosophical Writings*. Cambridge Texts in the History of Philosophy. Andrew Janiak (ed.), Cambridge: Cambridge University Press.

Nounou A, Psillos S (2012) Review of De Regt, Leonelli & Eigner (eds) Scientific Understanding: Philosophical Perspectives. *Studies in History and Philosophy of Modern Physics* 43: 72–74.

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

Psillos S (2009) *Knowing the Structure of Nature*. London: Palgrave-MacMillan.

Railton P (1981) Probability, Explanation and Information. *Synthese* 48: 233–56—reprinted in David-Hillel Ruben (ed.) *Explanation*, (1993) Oxford: Oxford University Press, pp. 160–81.

Rankine W J M (1855) Outlines of the Science of Energetics. In *Proceedings of the Philosophical Society of Glasgow* 3, eds Jürgen Renn et al., 2007. *The Genesis of General Relativity*, 4 volume-set. (Boston Studies in the Philosophy of Science, vol. 250.) Dordrecht: Springer

Renn J, Suer T (2007) Pathways out of Classical Physics. In *The Genesis of General Relativity, eds* Jürgen Renn et al. Dordrecht: Springer.

Salmon W (1998) *Causality and Explanation*, Oxford: Oxford University Press.

Smith G E (2014) Closing the Loop: Testing Newtonian Gravity, Then and Now. In *Newton and Empiricism*, eds Zvi Beiner and Eric Schliesser, 262–351. Oxford: Oxford University Press.

Strevens M (2013) No understanding without explanation. *Studies in History and Philosophy of Science* 44: 510–515.

Thomson W (1987) *Kelvin's Baltimore Lectures and Modern TheoreticalPhysics*. Edited by R. Kargon and P. Achinstein. Cambridge, Mass.: MIT Press.

Invited paper

EXPLANATION, MULTIPLE PERSPECTIVES, AND UNDERSTANDING

Melinda Bonnie Fagan
Department of Philosophy
University of Utah, USA
e-mail: mel.fagan@utah.edu

Abstract:
Science is increasingly interdisciplinary, as evidenced by empirical measures, funding initiatives, and the rise of integrative fields such as systems biology and cognitive neuroscience. In this paper, I motivate and outline an account of explanation for interdisciplinary contexts, building on recent debates about scientific perspectivism. Insights from these debates yield an inclusive list of relations between models constructed from different perspectives, which I then refine and generalize into a simple taxonomy. Within this taxonomy of relations among models, I identify the set of relations applicable to interdisciplinary contexts, discuss concepts of unification associated with each, and introduce three further constraints which furnish norms for this variety of explanation. Finally, I discuss implications of this account for a recent debate about understanding and explanation. One important consequence of my view is that explanation in interdisciplinary contexts and understanding of individual agents in those contexts are not equivalent.

Key words: explanation, interdisciplinarity, perspectivism, unification, models.

1. Introduction

The aim of this paper is to motivate and sketch an account of explanation suited to interdisciplinary contexts. There are already a number of philosophical accounts of explanation on offer, the most well-known being deductive-nomological (D-N), causal, unificationist, and pragmatic.[1] Each of these makes different assumptions about what explanation is, and what bears explanatory weight: causal powers or regularities, fundamental processes, general laws, counterfactual dependencies that give us control over natural processes, principles that unify disparate phenomena, simplified "minimal models" that identify core features of complex systems, and more. The D-N account, for example, states that to explain an event is to derive a description of its occurrence from a true general law and initial conditions (Hempel 1965); the pragmatic theory states that to explain is to provide an answer to a why-question (van Fraassen 1980); and interventionism equates explanation with revealing the factors on which the values of variables of interest depend (Woodward 2003). Causal theories of explanation range from the mechanistic account for life and behavioral sciences (Bechtel and Abrahamsen 2005, Craver 2007) to exchange of conserved quantities in physics (Salmon 1984).

None of these proposed accounts of explanation has proved universally satisfactory. Rather, the different accounts conform to and illuminate practices in different areas of science.[2] This is unsurprising, as philosophers formulate their accounts of explanation using a few key scientific exemplars—traditionally from physics, but increasingly from biological, behavioral, and social sciences as well. A plurality of

[1] Minimal model, kairetic, and topological explanation might also be included. See Woody (2015) for a critical overview of the major theories of explanation, and an alternative, "functional account" which has significant parallels to the view proposed here.
[2] Van Fraassen's pragmatic view is an exception, but that theory places few constraints on what can count as an explanation.

accounts of explanation is a consequence of variation in aims and standards of explanation across the sciences; there are multiple kinds of scientific explanation, which answer to different purposes and contexts. The multiplicity of philosophical accounts of explanation, mirroring diversity among the sciences, is for the most part unproblematic—in this sense, I am a pluralist about explanation. But the plurality of views about explanation can be problematic in interdisciplinary contexts, where fields committed to different views of explanation must work together. Such cases are increasingly common, given trends toward scientific specialization on the one hand, and complex real-world applications on the other. So, the multiplicity of accounts of explanation already on offer notwithstanding, it is desirable to have an account tailored for interdisciplinary contexts in particular.

This paper uses recent debates about scientific perspectivism to clarify a notion of explanation for such contexts. The next section provides background on interdisciplinarity in scientific practice, and the specific example that motivates my project. Section 3 summarizes the recent debate over scientific perspectivism, outlining the main positions and their insights for my purposes. Section 4 "re-purposes" insights from discussions of perspectivism into a framework for articulating a new account of explanation for interdisciplinary contexts. To preview the main result, the perspectivism debate affords a (partial) taxonomy of relations between models from different perspectives. The next task is to identify the family of relations within this larger system that are explanatory in the sense I am interested in. I do this by proposing what I take to be plausible constraints on these relations. The result is a sketch of an account of explanation that is suited to interdisciplinary contexts. Section 6 discusses implications of this account for recent debates about understanding, showing that explanation in interdisciplinary contexts is distinct from individual agents' understanding in light of such explanation.

2. Interdisciplinary contexts

The motivation for this new account of explanation is the prevalence of interdisciplinary contexts in scientific practice. Although there is no single agreed-upon definition of "interdisciplinary" or of "discipline," the broad sense in which I use them is well-established in the scholarly literature on interdisciplinarity.[3] By an *interdisciplinary context*, I mean a research situation that is proposed, planned, initiated, and carried out (or any subset of these) so as to require contributions from more than one discipline. I interpret "discipline" as including not only the traditional scientific and scholarly disciplines (e.g., biology, chemistry, physics, philosophy), but also fields and subspecialties such as "low temperature physics, molecular biology, developmental psychology, toxicology, operations research, and fluid mechanics" (Porter and Rafols 2009, 720). For my purposes, then, a *discipline* is a community of experts sharing a specialization with distinctive patterns of communication, methods, and standards of inquiry, whatever institutional entities those communities map onto (traditional disciplines, subdisciplines, specialties, and so on). In this I follow a number of social scientists who characterize as interdisciplinary "a genus of integrative research activities that combine more than one discipline, field, or body of knowledge" (Huutoniemi 2010, 309; see also Porter and Rafols 2009, quoted above, and Weingart 2010, 11). My usage is also consonant with that of Darden and Maull (1977), who refer interchangeably to disciplines, subdisciplines, and fields.

Obviously, not all research situations are such as to require contributions from multiple disciplines (broadly construed). Yet interdisciplinary contexts are prevalent enough to merit philosophical consideration in their own right. Funding for interdisciplinary research is a major theme of national and international science policy. Examples of funding organizations that explicitly encourage interdisciplinary research include the German National science program (DFG 1996, 202), the US National Science Foundation (Workshop report, 2008), the US National Academy of Sciences (2006), and the European Union's Horizon 2020 program (Horizon 2020). There is also empirical evidence that science has become increasingly interdisciplinary since the 1970s. Bibliometric analyses show a striking increase—more than 50%—in prevalence and complexity of co-authorship of scientific publications, as

[3] Mäki (2016) defines interdisciplinarity more broadly still: "whatever relevant relationship between two or more disciplines or their parts" (331). This definition includes comparisons of different disciplines as a species of interdisciplinary relation, while the notion of interdisciplinary context assumed here requires interaction of multiple disciplines in practice (including planning stages).

well as diversity of cited sources in article references, between 1975 and 2005 (Porter and Rafols 2009).[4] Social scientists Brint et al (2009) found steady growth of interdisciplinary programs in 4-year US colleges from 1975 to 2000, and the American Academy of Arts and Sciences charted a similarly steady increase in interdisciplinary PhD dissertations in the following decade (2003-2012; Figure 1). Although these measures do not capture the essential features of disciplines and interdisciplinary contexts as defined above, they do at least suggest that these contexts are increasingly important features of scientific practice. A more principled argument for the growing significance of interdisciplinary contexts considers the implications of "hyper-specialization" (Millgram 2015). Scientific inquiry (and inquiry more generally) is increasingly specialized, with ever-narrower fields of expertise propagating and dividing. Yet the goal of applying scientific knowledge to real-world problems persists, and has arguably even intensified in the past decades (Koskinen and Mäki 2016). So it is unlikely that any single discipline will be able to solve problems that do not originate within its own narrow framework. Interdisciplinary contexts would seem to be the only way forward, an unavoidable consequence of increased specialization and the demand that scientific knowledge be applicable beyond the specialized communities that produce it.

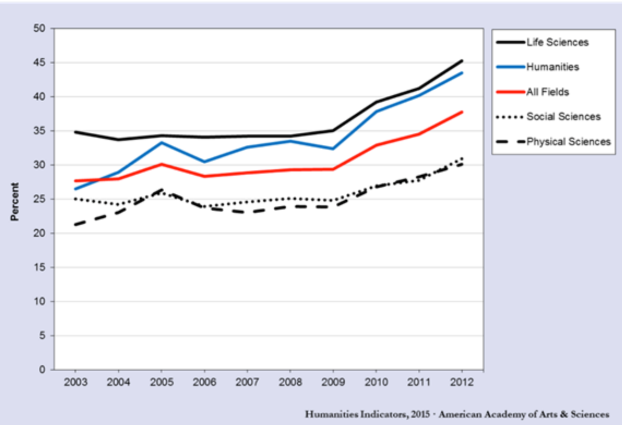

Figure 1. Empirical evidence of increased interdisciplinarity, 2003–2012

This paper is concerned with the class of interdisciplinary contexts in which the contributions of multiple disciplines are aimed at constructing an *explanatory model*. Of course, not all interdisciplinary contexts involve such aims, but there is good reason to think that many do. First, the concept of interdisciplinarity is often associated with integrated or unified theories or models. J.T. Klein, a prominent scholar of interdisciplinarity, traces the origins of the concept to "ideas of a unified science, general knowledge, synthesis and the integration of knowledge" (1990, 19).[5] Explanation is a traditional task of general, unifying theories. Similarly, the epistemic benefits commonly attributed to interdisciplinary research—theoretical innovation, reliable and accurate prediction—are also functions of explanatory models (Mäki 2016, 331). Finally, explanation is implicated in the main practical benefit attributed to interdisciplinarity: solving complex real-world problems. A necessary condition for solution is "an adequate understanding" of the problem:

> In order to create a functional solution to a complex problem, the problem has to be *understood comprehensively*, and its attempted solutions should reflect the interests of all relevant stakeholders and employ all the available knowledge from different sources (Koskinen and Mäki 2016, 429; italics mine).

Providing understanding is another traditional task of explanatory models (discussed in Section 6). Putting these points together: the benefits of interdisciplinary research are commonly characterized as being realized by or requiring explanatory models.

These hoped-for advantages notwithstanding, and despite prominence of and support for interdisciplinary research programs, there is evidence that interdisciplinary research "underperforms" in practice. Although Porter and Rafols (2009) found a 30-year "sustained and important increase" in interdisciplinarity by bibliometric criteria, most of these connections were between neighboring fields. When their measurements were calibrated to reflect "disparity" between fields, the bibiliometric data "show a more modest growth" in interdisciplinary research (mostly around 5%)" (ibid, 721). The social scientists concluded that "science is indeed becoming more interdisciplinary, but in small steps…only modestly increasing the connections to distant cog-

[4] Multiple aspects measured: "the number of disciplines cited (variety), the distribution of citations among disciplines (balance), and, crucially, how similar or dissimilar these categories
are (disparity)" (Porter and Rafols 2009, 721).

[5] But see Grüne-Yanoff (2016) for critique of the assumption that interdisciplinary research is necessarily integrative.

nitive areas" (ibid, 719). These results (and some assessments of the results of funding initiatives so far) suggest that there are obstacles to interdisciplinary research.

Furthermore, some of these obstacles concern explanation. In previous work, I have examined a "failure to communicate" between two communities each aiming to explain the regenerative capacities of stem cells (Fagan 2016). Stem cell biologists use experimental methods to characterize the molecular and cellular mechanisms underlying those regenerative capacities, and their effects on tissues, organs, and whole animals. Theoretical systems biologists, in contrast, construct simple dynamical systems models of cellular development, defining stem cells as "metastable local attractors on epigenetic landscapes." These systems biologists have repeatedly tried to bring their work to the attention of stem cell biologists, with little success. This failure of interdisciplinarity, I argue, is due to the two communities' different views of explanation. The systems modelers are committed to something like the D-N account, while most stem cell biologists aim at detailed mechanistic explanations (see Section 1). Interdisciplinary collaboration is blocked because each community assumes a view of explanation that is incompatible with the other: specific mechanistic detail versus abstract principles covering many cases. This case is a "proof of principle" showing that accounts of explanation grounded in one scientific discipline or another can impede progress in interdisciplinary contexts.

Furthermore, this problem is not limited to the stem cell case. Many biologists working in molecular/cellular/developmental biology are committed to something like the mechanistic model of explanation, and many dismiss systems biology as irrelevant to their projects.[6] But that attitude threatens the core idea of systems biology, which is to integrate experimental biology with other fields, such as physics, computer science, and engineering (Kitano 2002, O'Malley and Dupré 2005, Klipp et al 2009, MacLeod and Nersessian 2015). The success of systems biology depends on interdisciplinary contexts yielding (among other products) satisfactory explanations.[7] This is also the case for nanotechnology, climate science, cognitive neuroscience, environmental studies, and other prominent areas of research today. Other examples of failed interdisciplinary explanation have been documented in synthetic biology/humanities (Rabinow and Bennett 2012) and prokaryote/eukaryote phylogeny (O'Malley 2013). It is reasonable to suppose that many interdisciplinary projects in these and other areas fail to thrive, or do not get started in the first place, hampering our ability to tackle complex real-world problems. This situation motivates a new account of explanation, tailored to interdisciplinary contexts. The following sections lay the groundwork for such an account.

3. Perspectivism

In developing this account, I use recent work on scientific perspectivism as a resource. A *perspective*, in the everyday sense of the term, is a stance from which an object appears (to one or more observers), roughly synonymous with a "point of view." From different perspectives, the same object may present very different, even incompatible, appearances.[8] A classic example of incompatible perspectives is the parable of blind men studying an elephant: one has access to the tail, another to the trunk, another to a leg, and so on (Figure 2). From his experience with

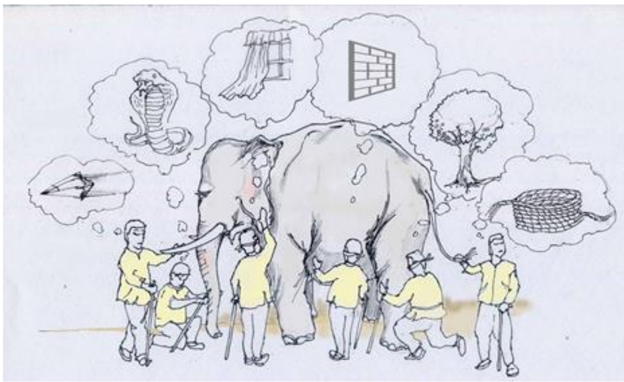

Figure 2 The blind men and the elephant, showing different perspectives based on partial experience of the object of inquiry. (Image from JainWorld.com)

one part of the elephant, each man conceives a representation of the whole animal: a rope, a snake, a tree, *etc*. Each representation is different, and, taken

[6] Personal communications (Stanford University, University of Utah).

[7] MacLeod and Nersessian (2015) characterize the interdisciplinary relation between (one strand of) systems biology and molecular biology. The explanatory aim they discuss is one-sided, however, involving only systems biologists.

[8] The term "perspective" also has a technical meaning in art (discussed in Giere 2006). I will not pursue that connection here.

as a representation of the whole, contradicts the others. None, individually, is a satisfactory representation of the object of inquiry. A satisfactory model of the elephant would integrate (or unify) the different partial representations conceived by the blind men, yielding the familiar image of the animal we know.

Tellingly, however, there are many different versions of this parable.[9] In many, the blind men never resolve their dispute – sometimes they continue to argue, or come to blows. In others, a "sighted man" (a king or a wise man, in the Jain version) conveniently passes by and explains "All of you are right. The reason every one of you is telling it differently is because each one of you touched the different part of the elephant. So, actually the elephant has all the features you mentioned."[10]

In philosophical discussions of scientific modeling, a *model* is taken to offer a (partial, often idealized) representation of a target object (*e.g.*, Giere 1988, Morgan and Morrison 2001, Weisberg 2013). Although there is not universal consensus on how to understand the representation relation linking models to targets, one very influential view is that representation by models implicates user's purposes – in other words, models represent their targets *from a perspective*. It is therefore unsurprising that recent discussions of perspectivism among philosophers of science were initiated by Ron Giere, who has long defended the view that representation of targets by models is a four-place relation involving—in addition to the model-target dyad—model-users and their purposes (Giere 1988, 2004). In *Scientific Perspectivism* (2006), Giere defends "perspectival realism" as a viable alternative to traditional scientific realism about theories, and as distinct from anti-realism/relativism. His argument has framed subsequent debate on these issues, which has focused on the compatibility of realism and multiple perspectives: can multiple partial models support realist conclusions about their target phenomena? To cast the debate in terms of the parable above, reality is the elephant in the story (so to speak), scientists committed to different models of (some part of) reality are the squabbling blind men, while traditional scientific realists take on the role of the helpful king. Perspectivists (including Giere) object that no human being can legitimately play the latter role.

Although the relation between models and the world is of primary concern in the perspectivism debate, the issue of how models from different perspectives relate *to one another* is also central. So there is a structural analogy between the perspectivism debate and my concern in this paper, explanation in interdisciplinary contexts. The different fields involved in cases of interdisciplinary research bring different perspectives to bear on the object of inquiry. Social scientists sometimes describe interdisciplinarity in exactly these terms. For example, Stichweh (1992) concludes his sociological study of scientific disciplines with the statement that: "Differentiation along disciplinary lines has the great advantage of viewing reality from radically different angles. The risks that the one-sidedness of each perspective entails are thus avoided" (12). In an important subclass of interdisciplinary contexts, the goal is to construct a single explanation from these diverse perspectives. For now, I wish to remain neutral as to whether explanations in interdisciplinary contexts must be understood realistically.[11] Bracketing that issue, perspectivists and I are concerned with different specifications of one general question: how can multiple diverse perspectives relate to one another so as to yield an epistemic product that is in some sense unified? For perspectival realists, the epistemic product is knowledge of the one real world; for myself it is a single explanation constructed in an interdisciplinary context. Below (Section 4), I argue that these are different senses of unification—another reason to treat the questions as analogous rather than coinciding.

In this and the next section, I repurpose several insights from the recent debate about perspectival realism in order to begin to articulate an account of explanation for interdisciplinary contexts. I first survey the various positions within the debate, and indicate the key insights taken from each.

As noted above, the recent debate begins with Giere's *Scientific Perspectivism* (2006). In this work, Giere proposes perspectivism based on the metaphor of color vision as a form of realism that avoids the excesses of traditional "objectivist" scientific realism without collapsing into constructivist anti-realism. His position is that scientific theories and observations (which for Giere are both models) are perspectival in the same way as descriptions of

[9] For an inclusive survey, see Wikipedia contributors (2016).

[10] "Elephant and the blind men." Jain Stories. JainWorld.com. Retrieved 2006-08-29.

[11] An affirmative answer would collapse the two issues that I present here as analogous. It would also contradict de Regt's (2015) account of understanding as not requiring truth. I will explore this issue in future work.

color: they tell us what the world appears to be like, from a particular standpoint. All justified realist claims by scientists are therefore "qualified" and "conditional," and should be interpreted as being of the form: "According to this highly confirmed theory (or reliable instrument), the world seems to be roughly such and such" (5-6). A consequence of Giere's perspectival realism is that quite different descriptions of the same bit of the world are not incompatible. Rather, each describes a different way of experiencing the world, such that one is not a better representation of that world than another. There is no real fact of the matter about which is correct. Disagreement (and normative selection of 'the best') is possible only within a perspective.[12]

Giere's account contrasts with objectivism, according to which, when models constructed from different perspectives represent the world in incompatible ways, at most one can be (approximately) true—there is direct conflict between models that must be resolved. As noted above, Giere has to a great extent set the terms of debate: is perspectivism a viable realist alternative to objectivism (traditional scientific realism) and to anti-realism? Or does it collapse into one or the other? In this way, objectivist assumptions frame the perspectivism debate. According to objectivism, when perspectives conflict (i.e., when models constructed from different perspectives represent the world in incompatible ways) at most one can be correct (or true). So the task is to select the perspective that is most likely correct, from a set of competing alternatives.[13] Giere rejects this objectivist assumption, but he does not say much about my primary concern: how models from different perspectives might relate to one another in other ways than direct conflict among competing alternatives.[14]

[12] A disanalogy: we know a lot about the mechanisms underlying different modes of color perception. Not just any visual perception set-up will work; there are many constraints on the experiences of dichromats and trichromats. Clearly it is not just "anything goes" in the case of color vision. But we do not have a corresponding sense of the constraints underlying the perspectives that produce scientific theories, models, and observations, so it is harder to rebut the "anything goes" objection.

[13] A pragmatic analog frames the task as selecting the perspective that is most useful for the problem at hand.

[14] Giere does claim that agents can switch perspectives and compare results among them (83-84), and his account of an experimental test as the "meshing" of theoretical and observational perspectives does involve a form of interaction between them (89, 93). But he does not examine these ideas in detail.

Chakravartty (2010) extends Giere's account in the direction of traditional realism.

His view is that seemingly incompatible models describe the actual dispositions of a real-world target, which are differentially revealed by different methods of detection; i.e., "from this perspective the system reveals disposition X, from that perspective it reveals disposition Y" (paraphrasing Rueger 2005, 581). So apparently incompatible models are reconciled by seeing them as revealing different dispositional facts about a common target system (410). Explanations, according to Chakravartty, either answer contrastive "what-questions," picking out the appropriate answer among the multiple perspectives on offer, or describe underlying (non-perspectival) facts about the system's structure that produce the different dispositions under different circumstances. Although it is not his main focus, Chakravartty does imply an account of relations among different models of the same target. The claims of different models of a target T can be united in a long conjunction stating each method and the dispositions of a target revealed by that method. Such a conjunction has the form: "Method x reveals a about T, and method y reveals b about T, and so on." For example, "What is water? 'It is something that dissolves salt in some circumstances and not in others; that behaves like a continuous medium in some circumstances and not in others; that is nourishing in some circumstances and not in others' and so on…" (ibid, 412). Hereafter, I will refer to this relation between models from different perspectives as *simple additivity*.

Morrison (2011) takes the opposite tack from Chakravartty, arguing that in some cases multiple inconsistent models (like those of atomic nuclear structure) are tantamount to anti-realism. In these cases, she argues, different models ascribe inconsistent "fundamental properties" to the target of inquiry, so "there is no way to build on and extend the models in a cumulative way" (351). This kind of incompatibility is conflict as the objectivist sees it—models for which there is no unifying core or coherent treatment of all the alternatives. In other cases (such as models of turbulence), Morrison sees different models as unproblematic. Such models, she argues, are different ways of elaborating the same set of basic principles—alternative idealizations from a shared common structure. Antirealist concerns are blocked by the availability of a unified treatment of the different models—a shared core set of basic principles. Morrison refers to this relation between models as "complementary," but (as I discuss below) it is more like Kitcher's explanatory unification (1981, 1989) or Mitchell's "theoretical unification" (2003). In what follows I refer to this relation as *subsumption*. To summarize: Morrison distinguishes two kinds of incompatibility among models of the same phenomenon: one compatible with theoretical unification (and traditional realism), the other supporting anti-realism. The former I term subsumption, while the latter coincides with the objectivist conflict. So Morrison rejects Giere's "third way" concerning realism.

Chirimuuta (2016) offers a different criticism of Giere, arguing that the metaphor of vision obscures some resources for perspectival realism, and that the sense of touch, the operation of which involves active engagement and interaction with the world, is a more appropriate guiding metaphor. She argues persuasively that the visual metaphor reinforces the objectivist assumption that scientific theories, models, and observations "should not be conditioned, or molded by the interactive process through which they have come about" (753). Vision seems to be a sense that works like this: the object presents itself to us, without our making an effort, bringing our own activity/subjectivity to bear on it. Yet our scientific knowledge about vision contradicts this impression. Focusing on the sense of touch is a good way to offload the baggage of the visual metaphor. Chirimuuta emphasizes that "scientific representations inform us about the natural world in virtue of their interactive and interested qualities" (746). What might appear, on the visual metaphor, to be distortion can be a means to truth, following Chirimuuta's "haptic metaphor." Although she does not discuss relations between different models, Chirimuuta's characterization of the model-world relation as an interactive process has important implications for inter-model relations as well. I return to this point in the next section.

Massimi (2016a, b) also emphasizes interaction, and an alternative to Giere's perspectival realism. Her key idea is success across perspectives, such that successful claims of scientific knowledge from a given perspective must also meet standards of adequacy when assessed from other perspectives. This is a perspectivalist account of "success-to-truth inferences" (i.e., realist inferences) "from within"—i.e., without presupposing an objectivist "view from nowhere," or elevating our own perspective to that of an arbiter. The crucial idea is a form of interaction among perspectives: "cross-perspectival assessment" (2016a, 12). Realism, Massimi argues, involves a historical prediction: approximately true insights will be preserved across theory change. A scientific claim (or model) meets Massimi's criterion of "success from within" just in case it performs adequately with respect to standards of its own, originating perspective, expresses a proposition that is in fact true (not that we are ever in a position to demonstrate this), and—crucially—also meets standards of performance-adequacy in its original context when it is assessed from other scientific perspectives. More precisely:

> "A scientific claim (SC) meets the criterion of success from within iff:
>
> (a) SC expresses a proposition p at scientific perspective SP1
>
> (b) p is true (i.e. corresponds to states of affairs in nature) and meets standards of performance-adequacy in SP1 when assessed from other scientific perspectives SP2, SP3, SP4 ..." (2016b, 765).

I take Massimi's point to be that the standards of performance-adequacy in SP1 (the original perspective in which a claim is made) have to make sense in other perspectives—that there is a way to translate the concepts involved in making the judgment in SP1 so that modelers in other perspectives can judge that model's adequacy by its own lights as those lights are understood in the other perspectives. If this is correct, then Massimi posits a one-sided interaction among perspectives, such that models are continually assessed for "ongoing performance-adequacy" in terms of the original standards as those standards appear in other perspectives (ibid, 767).[15]

4. Taxonomy of relations

The recent debate over perspectival realism yields a list of relations that can obtain between models from different perspectives:

- direct conflict
- simple additivity
- subsumption
- interactive process
- cross-perspective translation
- no cumulative interaction

[15] Full quotation: "Each scientific perspective—I suggest—functions then both as a context of use (for its own knowledge claims) and as a context of assessments (for evaluating the ongoing performance-adequacy of knowledge claims of other scientific perspectives)" (767).

• non-interaction

The first five are just the relations revealed in the literature survey above. Direct conflict is the objectivist view, such that we select the best model among alternatives. Simple additivity is the bare conjunction of statements from different models associated with different methods. Subsumption by more basic principles (traditional unification) is an indirect relation between models—each is subsumed by the shared basic principles. The idea of an interactive process is not a relation per se, but part of a more general framework for thinking about these relations (see below). Cross-perspective translation is a kind of interactive process, but with distinctive features; I return to this point shortly. Lack of cumulative interaction is an issue raised by Morrison (2011); I return to this point below also. Non-interaction is a limit case: absence of a relation between models in practice. These seven possibilities, loosely articulated as they are, present a more nuanced set of alternatives than the standard typology of inter-model relations: insulation (what I have termed "non-interaction"), interaction, and integration into "relatively coherent systems" (Koskinen and Mäki 2016, 424).

The next task is to narrow down this list, in order to pinpoint the relations applicable to explanation in interdisciplinary contexts. Such contexts include two or more perspectives, all of which are needed to construct an explanatory model of the phenomenon of interest.[16] So a few of the listed relations can be eliminated as inapplicable to such contexts: non-interaction, lack of cumulative interaction, and direct conflict. The last is, to be sure, a form of interaction, but not one that is helpful for understanding interdisciplinary contexts, where (I assume) relations among contributing models must be integrative in some sense. The task is not to select the best perspective for the purpose at hand, where "the best" is assessed from a stance outside all the different perspectives—the "view from nowhere" that perspectivists repudiate. The challenge is rather to articulate, as Massimi puts it, a stance "from within" which connects the different perspectives.

Simple additivity, I think, can also be ruled out. Most if not all interdisciplinary contexts in which concerns about explanatory models arise will not fit this pattern. The reason is that simple additivity involves no substantive connection between models from different perspectives; their contributions are simply strung together in a conjunction. The "unification" of simple additivity is bare logical consistency. There will of course be many cases in which relations among models from different perspectives do fit this pattern, but these contexts will not be helped by the account of explanation that I want to propose. This insight reveals a requirement for this account of explanation: diverse models contributing to explanation in an interdisciplinary context should be unified in a sense stronger (or more substantive) than the unity of a consistent conjunction. I will refer to this as the *unification requirement* for explanation in interdisciplinary contexts.

Finally, an additional relation between models should be added:

• complementarity

Morrison (2011) refers to theoretically unified models—disparate idealizations of a shared fundamental model—as "complementary." I have instead used the term "subsumption" to describe this relation, albeit indirectly. But complementarity is a familiar relation among different models. It is conceivable (and in fact more consistent with ordinary usage) that models from different perspectives can complement one another through differences that do not converge on a common core. I will say more about this relation shortly; for now, I merely add it to the list:

• subsumption
• interactive process
• cross-perspective translation
• complementarity

These four remaining relations can, with some rearrangement and generalization, be arranged in a taxonomy. This taxonomy is framed by two cross-cutting distinctions: same-different, and process-given (Table 1). The "process-given" distinction concerns how we conceptualize the relation between models from different perspectives. One option is to treat models from different perspectives as completed constructs that need to be connected after the fact (the different models are "given"). Or, following Chirimuuta's haptic metaphor, we can think of perspectival integration as a process occurring over time, as models are constructed. The "same-different" distinction concerns the features of mod-

[16] I will assume that all models target the same phenomenon. The implications of relaxing this assumption will be explored in future work. The motivating concerns about interdisciplinary contexts (Section 2) suffice to justify the assumption of a single target at this stage of the project.

els in virtue of which they connect with one another. Models could connect through sameness: having features common to both. Or they could relate to one another through difference: a contrast leveraged into a positive link. These are not incompatible alternatives; most models from distinct perspectives will share some features and not others, as finished products or during construction. So relations relevant to explanation in interdisciplinary contexts could involve both; the same/different distinction is not a dichotomy. For simplicity, however, I will discuss the kinds of relations revealed by this taxonomy in four categories.

Table 1. Taxonomy of relations between models from different perspectives, applicable to interdisciplinary contexts

	process	**given**
same	cross-assessment	subsumption overlap
different	complementary	contrastive

The subsumption relation that Morrison discusses is located in the "same-given" quadrant, for obvious reasons. But it is not alone in that quadrant. Subsumption, as noted above, is not of one perspective by another, but of both—a more "fundamental" model of their common features. Another possibility is to reject that hierarchical assumption, and think of the shared features as "overlap" between models; i.e., models from distinct perspectives coincide in certain respects, and these areas of coincidence serve as "joints" that connect them. Massimi's cross-perspective assessment is located in the "process/same" quadrant. It too is only one example, as there are many other possible ways to achieve "sameness" in models from different perspectives as an interactive process. Models vary so much in their features and resources for construction that a more substantive general statement concerning this quadrant is probably not possible.[17] To return to Massimi's example: cross-perspective assessment is a process—specifically, a historical process. Massimi considers perspectives in historical lineages—changes within a continuous strand of research over time; for example, theories of light. For these cases there is a kind of "inheritance with modification" of concepts and modeling strategies. The perspectives Massimi considers are related as ancestor and descendants. It is in the "same" category because Massimi's notion of cross-perspectival assessment requires enough overlap in concepts and aims for an assessment of one in terms of the other to be possible.

There is an interesting contrast to interdisciplinary contexts here. Some degree of cross-perspective understanding does seem necessary for interdisciplinary projects to succeed. But it seems too much to demand that standards "translate" from one perspective to another, as Massimi's account requires. Perspectives that are separated by disciplinary boundaries are not typically related as ancestor-to-descendants (although this might happen in some cases, such as synthetic and molecular biology). Although perspectives involved in interdisciplinary contexts might have a common ancestor, they are contemporaneous and have distinct social and scientific identities.[18] The "same-process" quadrant allows for many kinds of similarity to play a role in model-construction.

Paralleling the quadrants in the "same" category are analogous generic characterizations of relations between models from different perspectives, such that the connection is achieved over the course of model-construction, or established between individually "given" models. But rather than relating to one another in virtue of shared features, models can be connected in virtue of their differences. The most familiar example of this is contrastive explanation, which reconciles apparently conflicting models of the same target by an explanatory division of labor. This example is located in the 'different-given' quadrant, because on the contrastive (erotetic) account, explanatory models answer distinct questions about the phenomenon of interest. The "different-process" quadrant is not exemplified by any positions in the perspectival realism debates. Nonetheless, relations in this category are conceivable—models in different perspectives are, at some point in their construction, fitted together like jigsaw puzzle-pieces. I therefore term this category of relation "complementary." Cross-checking and mutual con-

[17] However, Nersessian's concept of a "problem-space" may be a useful framework for categorizing the diverse tools for model-construction.

[18] The biological counterpart would be the interplay of homology and analogy in biology—similarity due to shared inheritance, vs. similarity due to convergence on the same adaptive "solution."

straint on possibility space are examples of relations in this quadrant.

To summarize this section so far: I have narrowed down and generalized the list of relations between models from different perspectives that emerged from consideration of recent debates over perspectival realism. The narrowing-down was accomplished by ruling out relations that are not plausibly applicable to interdisciplinary contexts. Generalization was achieved by locating the remaining relations in a taxonomy organized by same/different and process/given distinctions (Table 1). The next step is to consider these four general categories of relation, and to build from these toward an account of explanation in interdisciplinary contexts. I noted earlier the requirement for explanation in interdisciplinary contexts that the models from different perspectives be unified in a sense stronger than simple additivity. The taxonomy based on insights from the perspectivism debate offers ideas about what these senses might be. I distinguish three forms of unification that emerge from the taxonomy above: top-down, bottom-up, and complementary.

The first of these is familiar, being akin to Kitcher's explanatory unification (1989). Kitcher defines explanation as the logical derivation of many facts via a minimal set of general argument patterns. Unifying explanations in Kitcher's sense are arguments—the premises of which must contain at least one empirical law of nature—and initial conditions specifying how it applies, such that the conclusion—a description of the event (or regularity) to be explained—logically follows. To explain an event or regularity, on this and the covering-law view, is to show that it is to be expected, subsuming it to the wider domain "covered" by the law. Kitcher's explanatory unification is achieved through a shared argument pattern, which plays the explanatory role assigned to laws in the traditional covering law (D-N) view. However, Kitcher's account is holistic, or global, in a way that is not presumed here: a unifying explanation in his sense fits one of the set of argument patterns that "collectively provide the best systematization of [all] our beliefs" in a particular socio-historical context (1989, 430).[19] Shorn of this global character, unification by a general argument pattern is a special case of what I will term "top-down unification." Applied to perspectives, unification of this sort is via similarities among models that can be expressed as an overarching third term: the common pattern that connects them. This is just the relation between models referred to above as "subsumption."

A second form of unification dispenses with the overarching third term, so that models from different perspectives are directly related by similarity. I refer to this "bottom-up unification" for contrast. The shared common ground may be pre-formed (given) in each model, or constructed in the pursuit of interdisciplinary goals (process). Winsberg's (2010) study of computer simulation includes examples of the latter. In Chapter 5, he describes how models constructed using different, incompatible theories can be ingeniously "sewn together" (made to "shake hands") to comprise a single model that is used to simulate propagating cracks in solids. The nature of the stitches (and whether they trade on sameness or difference) varies, even within the three perspectives he considers. Winsberg emphasizes that some connections among models are "useful fictions" that allow models constructed from different theoretical perspectives to "meet at the edges." Insofar as this stitching together of different models is accomplished via features that the models share (as given or after an interactive process), the relation is what I term "bottom-up unification."

Lastly, there is unification of dissimilar parts via complementarity. This is closer to the everyday meaning of the term: "combining two or more things to or with one another so as to form one whole or unit." [20] The additional qualifier here is that the things that combine (e.g., models from different perspectives, resources for model-construction) do so in virtue of their differences rather than similarities. Unification in this sense is constructive, such that a more inclusive whole is formed out of diverse elements; for example, covalent bonding between molecules with complementary shape and charge. The disparate things, in the class of cases at issue in this paper, are (aspects of) models from different perspectives. This form of unification also involves a "third thing"—a larger whole constituted by disparate parts. The different perspectives complement one another; their differences are a resource rather than cause for epistemic concern. Scientific examples of complementary unification are common; e.g.,

[19] A significant feature of Kitcher's account, which it shares with many other theories of explanation, is that explanatory power is associated with *reducing diversity*. An explanation, for Kitcher, identifies a shared common pattern in each of a set of diverse things: facts, beliefs, or phenomena.

[20] Paraphrase of the *OED online* definition of "unify" (accessed Aug 2015).

multi-level explanations that relate phenomena at different levels of biological organization through the details of particular experimental results.

5. Constraints and norms

To further clarify explanation for interdisciplinary contexts, I propose three further constraints on these forms of unification. These constraints together provide a sketch of an account of explanation for interdisciplinary contexts.

5.1 *Pragmatism*

The first constraint is pragmatic: that a unifying relation linking models across perspectives does not immediately present new explanatory challenges from within any of those perspectives. That is, from the point of view of the perspectives so connected, the connecting relations are not immediately pressing topics of inquiry in their own right. If this condition is not satisfied, the cross-perspectival explanatory model would depend on a relation that itself cries out for explanation by the standards of one of the involved perspectives. Connections ("stitches") across perspectives that yield unified explanatory models in the sense at issue allow us to see "the forest *and* the trees," so to speak—they extend our view to encompass both perspectives, without presenting us with further puzzles as to how forest and trees are related.

5.2 *Mutuality*

The second constraint is an extension of Massimi's cross-perspectival assessment. An explanation comprised of multiple models must intelligibly relate those models' different viewpoints. That is, the relation between perspectives—the way in which they are unified—is assessed as successful according to the standards of all perspectives thereby connected. So the unifying relation that integrates models from distinct perspectives must be one that meets the standards of all the involved perspectives. That relation need not play *the same role* in each perspective. But it must have *a role* in each perspective. This ensures that the unifying relation is *mutual* across the relevant perspectives. This is quite different from requiring that each contributing model be acceptable to all involved perspectives—although such cases are not ruled out. Rather, the requirement concerns the unifying relation (top-down, bottom-up, or complementary) that connects models from different perspectives.

This condition, I think, offers a way forward for the systems biology example in Section 2. To review, the problem in that example is that each community assumes a view of explanation that aims at virtues incompatible with the other: specific mechanistic detail (for stem cell biologists) versus abstract principles covering many cases (for systems modelers). The positive solution I proposed in Fagan (2016) was that, for dialectical reasons, the systems modelers should adopt a mechanistic framework and accommodate their models to it. This is a one-sided solution, which reinforces existing power-relations within science. An account of explanation for interdisciplinary contexts (or others) should be independent of those factors. The mutuality condition allows for a more incisive critique of the case. The problem with systems modelers' proposed explanations of stem cell behavior (and of cell development more generally) is not that they appeal to general principles and not detailed mechanisms, but that their "image" of how systems models bear on experimental biology is not acceptable by the standards of stem cell biology (and many other life science fields). The reverse is also true; hence the need for a new account of explanation that requires this kind of mutuality.

5.3 *Comprehensiveness*

A third constraint comes from an important explanatory virtue: comprehensiveness.[21] This virtue is contrasted with narrow, parochial viewpoints on a phenomenon of interest. To return to the "blind men and the elephant" example: while each man has a model that is correct, as far as it goes (i.e., to the part of the elephant that is similar to that representation), the parable depends on the intuition that the clash of incompatible models is unsatisfactory. The problem, though, is not only that the models are inconsistent, but that each is woefully incomplete. The blind men's models of the elephant lack comprehensiveness. Of course, no model is perfectly or ideally comprehensive (e.g., Teller 2001). But as a comparative notion, comprehensiveness is a useful criterion for evaluating relations between models from different perspectives. One may ask, in a given case, does the unifying relation between models from different perspectives yield an integrated model that

[21] Chirimuuta hints at this with a remark: "to achieve a more comprehensive view of the object one may stitch together an array of partial representations" (10). But the haptic metaphor she proposes doesn't suggest how this might work.

is more comprehensive—that allows us to *at once* engage with more aspects of the phenomenon of interest—than the models of isolated perspectives? Such a model allows us to, so to speak, see the forest *and* the trees, rather than forcing us to choose between one or the other. Comprehensiveness is undercut by a lack of unification among the models; if multiplying perspectives yields a heap of unrelated "snapshots" of a phenomenon, then increased comprehensiveness is not much of a gain. Unifying relations should, in general, increase the comprehensiveness of explanatory models.

Not all the relations I've discussed satisfy this condition. The aim of Winsberg's "handshaking model," for example, is not to explain, but to predict how materials will behave under strain. The integrated parallel multiscale model in this case is used to generate simulation results. But this seems distinct from the ability to comprehend multiple aspects of a phenomenon for ourselves. More speculatively, unification of perspectives, if unaccompanied by some "higher-level unity," is not explanatory (in the sense at issue). This suggests that "bottom-up" unification will not in general yield explanations for interdisciplinary contexts.

To sum up, I propose three constraints on unifying relations between models from different perspectives:

(1) **pragmatism:** the unifying relation does not immediately present new explanatory challenges from within any of the perspectives thereby unified

(2) **mutuality**: the unifying relation is assessed as successful by the standards of all perspectives thereby unified

(3) **comprehensiveness:** the unifying relation yields a model that is more comprehensive than the models of isolated perspectives

These three constraints correspond to norms for explanation in interdisciplinary contexts.
All these requirements apply to the unifying relations among models from different perspectives. Importantly, there is no requirement for an overall unified model—except in the minimal sense that models of multiple perspectives are connected—so it is possible to "comprehend" both representations at once. But what this amounts to may differ across the relevant perspectives. I conclude by relating these results to recent work on understanding.

6. Understanding in interdisciplinary contexts

One of the main debates among philosophers of science concerned with understanding is whether or not the topic is exhausted by philosophical accounts of explanation. The traditional view, influentially put forward by Hempel (1965) and recently defended (albeit with caveats) by Khalifa (2012, forthcoming), is that scientific explanations provide understanding, such that further philosophical study of understanding (beyond philosophical accounts of explanation) is redundant. This traditional view has recently been challenged, notably by de Regt (2009, 2015), who argues that "pragmatic understanding" is necessary for constructing explanations. Pragmatic understanding consists in the ability to use relevant theory to construct models of phenomena; theories are intelligible to the extent that they are useful in this role. So pragmatic understanding is based on the skills and judgments of scientists, which vary across communities (2009, 596). Therefore, de Regt argues, an adequate account of explanation requires an account of at least this form of understanding, which is distinct from, but required for, successful explanation of phenomena. As noted above, Khalifa rebuts de Regt's position, defending the "Explanatory Model of Understanding." More precisely, Khalifa poses a challenge to philosophers of understanding, to articulate ways in which understanding is distinct from explanation. My account of explanation in interdisciplinary contexts takes up Khalifa's challenge. I will show that, despite substantial agreement with Khalifa on scientific explanation, the account sketched in this paper drives a wedge between explanation and understanding.

The three norms presented above correspond closely to Khalifa's EKS Model of understanding (forthcoming, 12):

"(EKS1) S_1 understands why *p* better than S_2 if and only if:

(A) *Ceteris paribus*, S_1 grasps *p's* explanatory nexus more completely than S_2; *or*

(B) *Ceteris paribus*, S_1*'s* grasp of *p's* explanatory nexus bears greater resemblance to scientific knowledge than S_2*'s*."

"EKS" refers to explanation, knowledge, and science, the key concepts implicated in Khalifa's account of understanding. EKS is underwritten by two principles, the Nexus Principle (A) and the Scientific Knowledge Principle (B). The former states that an agent's understanding is improved by "grasping" more explanatory information relevant to explanandum *p*; i.e., more factors that explain *p*, as well as how those factors "hang together" (ibid, 5). Khalifa's "grasping more of p's explanatory nexus" is very similar to the comprehensiveness constraint presented above (Section 5.3). Both are comparative notions, and both depend on the assumption that including more information about a phenomenon is an explanatory virtue.[22] The Scientific Knowledge Principle states that an agent's understanding is improved by her belief that *q* explains *p* being "safe" (not likely to be false) due to a belief-forming process that more resembles "scientific explanatory evaluation" (ibid, 10). Khalifa's account of scientific explanatory evaluation is universal or monistic, while my account appeals to local standards of participating disciplines. But the pragmatic and mutuality constraints for explanation in interdisciplinary contexts amount to a counterpart of Khalifa's second principle.

At this point, however, an important contrast between Khalifa's account and my own emerges. The requirements for pragmatism and mutuality concern connections between models from different perspectives: the ways in which these different models of a phenomenon "hang together," to use Khalifa's term. *Within* disciplines, Khalifa's account may hold: explanations confer understanding, such that "[a]ny philosophically relevant ideas about scientific understanding can be replaced by philosophical ideas about scientific explanation without loss" (Khalifa 2012, 18). But my account of explanation in interdisciplinary contexts opens a wedge between explanation and understanding, which has no counterpart in the EKS Model. The wedge begins with the assumption that understanding is a cognitive attitude of individual human agents. All the various accounts of understanding, to my knowledge, share this assumption, forcefully articulated by de Regt (2009): "One can use the term 'understanding' only with—implicit or explicit—reference to human agents: scientist S understands phenomenon P by means of explanation E" (586). Apart from rare cases, the individual agents involved in interdisciplinary contexts will belong to just one of the contributing disciplines. An explanation that satisfies norms of pragmatism, mutuality, and comprehensiveness does *not* thereby confer "interdisciplinary understanding" on individual agents from the several contributing disciplines. This raises the question: what understanding for individual agents is conferred by an explanation of the sort I have motivated and sketched in this paper?

For individual agents, the gain in understanding is *relational*. An individual researcher S grasps, from a model unified in one of the three ways discussed above (top-down, bottom-up, or complementary):

> (1) the parts of the model constructed with resources from S's own discipline, and
> (2) the unifying relation that connects those parts with those constructed with
> resources from other disciplines than S's.

The explanatory model as a whole is comprehensive, yet need not confer comprehensive understanding on individual researchers. Instead, individual researchers understand both the features of the explanatory model that are grounded in their own discipline, and the connections to features grounded in other disciplines. They need not (and in most cases will not) understand the explanatory model as a whole. Yet the gain in understanding conferred by an interdisciplinary explanation is not trivial. Understanding of the unifying connections between models from different disciplinary perspectives gives individual researchers a path or bridge into a different disciplinary perspective. They do not thereby understand the explanatory model as a researcher from another discipline would, but they can see "how to get there from here." Even if they never cross that bridge, a counterfactual statement about that "road not taken" will be true:

> For an explanatory model M constructed using resources from disciplines D_1, D_2,…D_n, if S_{D1} were to become proficient in the resources of disciplines D_2…D_n, then S would understand M.

Knowledge of this counterfactual gives individual researchers "social epistemic understanding" of how their own expertise is related to that of other researchers.

Another way to put the point is that models constructed within particular disciplines—which confer understanding of their target on individual researchers—are unavoidably partial; i.e., such models are

[22] One contrast is that Khalifa's account requires that explanatory factors grasped be "correct;" i.e., approximately true and empirically adequate. As discussed above, my account of explanation in interdisciplinary contexts (for now) eschews such a requirement, allowing the standards of participating fields to determine this on a case-by-case basis.

incomplete, non-comprehensive representations of the target phenomenon. *Modelers* from particular disciplines are also unavoidably partial, in another sense of that term: modelers are biased in favor of the standards of their own disciplines, and will, *ceteris paribus*, tend to evaluate others by those standards. Explanations for interdisciplinary contexts undercut modelers' partiality through the social epistemic understanding they confer on individual researchers: they see their own discipline related to others through the unifying connections between different models. Insofar as the argument of this paper is successful, understanding and explanation are distinct, though related. Examining the relation between them offers a fruitful approach to the social epistemology of scientific inquiry.

7. Conclusion

The overall aim of this project is to develop an account of explanation that is suited to interdisciplinary contexts, rather than to particular scientific disciplines (conceived in the broad sense). I began by motivating this new account of explanation alongside those grounded in particular scientific disciplines. There is a perceived need for interdisciplinary research, and interdisciplinarity is a prominent feature of current science. At least in some cases, however, interdisciplinary projects underperform, for reasons having to do with explanation. It is therefore desirable to articulate an account of explanation suited to interdisciplinary contexts. To do this, I use recent debates about perspectival realism to construct a general taxonomy of relations between models from different perspectives. I then identify within this taxonomy the set of relations applicable to interdisciplinary contexts, discuss concepts of unification associated with each, and introduce three further constraints: pragmatism, mutuality, and comprehensiveness. These amount to a sketch of the new account.

In the final part of the paper, I discuss implications of this account for recent debates about understanding and explanation. I show that, on the account of explanation proposed here, individual researchers' understanding is distinct from explanation by a unified model constructed using contributions from multiple disciplines. Although such an explanation does afford improved understanding for participating researchers, the understanding they gain is of connections among their various fields, not necessarily of the explanatory model as a whole. Further work is needed to develop this account of explanation for interdisciplinary contexts, and the social epistemic relation between explanation of this kind and understanding.

Acknowledgments

Thanks to Lilia Gurova for the kind invitation to contribute this keynote paper to the special issue of the *Balkan Journal of Philosophy* on "Explanation and Understanding." An earlier version of this paper was presented at the University of Pittsburgh (November 2016); I thank the participants at that event for their comments and questions. This paper has also benefited from discussions with Hanne Andersen, Anjan Chakravartty, Mazviita Chirimuuta, Haixin Dang, Steve Downes, Kareem Khalifa, Jim Lennox, Michela Massimi, and Sandra Mitchell. Funding was provided by the University of Utah College of Humanities, and a generous donation from the family of Sterling M. McMurrin.

References

Bechtel W, Abrahamsen A (2005) Explanation: a mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–41.

Brint S G, Turk-Bicakci L, Proctor K, Murphy S P (2009) Expanding the social frame of knowledge: interdisciplinary, degree-granting fields in American colleges and universities, 1975–2000. *The Review of Higher Education* (Winter 2009) 32: 155–183.

Chakravartty A (2010) Perspectivism, inconsistent models, and contrastive explanation. *Studies in History and Philosophy of Science* 41: 405–412.

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

Chirimuuta M (2016) Vision, perspectivism, and haptic realism. *Philosophy of Science* 83: 746–756.

Craver C (2007) *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press, Oxford.

Darden L, Maull N (1977) Interfield theories. *Philosophy of Science* 44: 43–64.

de Regt H (2009) The epistemic value of understanding. *Philosophy of Science* 76: 585–597.

de Regt H (2015) Scientific understanding: truth or dare? *Synthese* 192: 3781–3797.

Fagan M B (2016) Stem cells and systems models: clashing views of explanation. *Synthese* 193: 873–907.

Frodeman R, Klein J T, Mitcham C (eds) (2010) *The Oxford Handbook of Interdisciplinarity*. Oxford University Press, Oxford.

Giere R (1988) *Explaining Science: A Cognitive Approach*. University of Chicago Press, Chicago.

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

Giere R (2006) *Scientific Perspectivism*. University of Chicago Press, Chicago.

Grimm S R (2006) Is understanding a species of knowledge? *British Journal for the Philosophy of Science* 57: 515–535.

Grüne-Yanoff T (2016) Interdisciplinary success without integration. *European Journal of Philosophy of Science* 6: 343–360.

Hempel C (1965) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press, New York.

Horizon 2020 (2014) The EU framework programme for research and innovation – Horizon 2020. Questions and answers.

Huutoniemi K (2010) Evaluating interdisciplinary research. In) *Oxford Handbook of Interdisciplinarity, eds* Frodeman R, Klein J T, Mitcham C, 309–320. Oxford: Oxford University Press.

Khalifa K (2012) Inaugurating understanding or repackaging explanation? *Philosophy of Science* 79: 15–37.

Khalifa K (forthcoming) *Understanding, explanation, and scientific knowledge*. Cambridge University Press, Cambridge.

Kitano H (2002) Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. *Current Genetics* 41: 1–10.

Kitcher P (1989) Explanatory unification and the causal structure of the world. In *Scientific Explanation, eds* Salmon W, Kitcher P, 410–506. Minneapolis: University of Minnesota Press.

Klein J T (1990) *Interdisciplinarity: History, Theory, and Practice*. Detroit: Wayne State University Press.

Klipp E, Liebermeister W, Wierling C, Kowald A, Lehrach H, Herwig R (2009) *Systems Biology: A Textbook*. Wiley-VCH, Weinheim.

Koskinen I, Mäki U (2016) Extra-academic transdisciplinarity and scientific pluralism: what might they learn from one another? *European Journal for Philosophy of Science* 6: 419–444.

MacLeod M, Nersessian N (2015) Modeling systems-level dynamics: understanding without mechanistic explanation in integrative systems biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 49: 1–11.

Mäki U (2016) Philosophy of interdisciplinarity. What? Why? How? *European Journal of Philosophy of Science* 6: 327–342.

Massimi M (2016) Three tales of scientific success. *Philosophy of Science* 83: 757–767.

Massimi M (forthcoming) Four kinds of perspectival truth. *Philosophy and Phenomenological Research*.

Millgram E (2015) *The Great Endarkenment: Philosophy for the Age of Specialization*. Oxford: Oxford University Press.

Mitchell S (2003) *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.

Morrison M (2011) One phenomenon, many models: Inconsistency and complementarity. *Studies in History and Philosophy of Science* 42: 342–351.

Morgan M, Morrison, M (eds.) (1999) *Models as Mediators*. Cambridge: Cambridge University Press.

O'Malley M (2013) When integration fails: Prokaryote phylogeny and the tree of life. *Studies in History and Philosophy of Biological and Biomedical Sciences* 44: 551–562.

O'Malley M, Dupré J (2005) Fundamental issues in systems biology. *BioEssays* 27: 1270–1276.

Porter A L, Rafols I (2009) Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics* 81: 719–745.

Rabinow P, Bennett G (2012) *Designing Human Practices: An Experiment with Synthetic Biology*. Chicago: University of Chicago Press.

Rueger A (2005) Perspectival models and theory unification. *British Journal for the Philosophy of Science* 56: 579–594.

Salmon W (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Stichweh R (1992) The sociology of scientific disciplines: on the genesis and stability of the disciplinary structure of modern science. *Science in Context* 5: 3–15.

Teller P (2001) Twilight of the perfect model model. *Erkenntnis* 55: 393–415.

Van Fraassen B (1980) *The Scientific Image*. Oxford: Oxford: Clarendon Press.

Weingart P (2010) A short history of knowledge formations. In *Oxford Handbook of Interdisciplinarity, eds* Frodeman R, Klein J T, Mitcham C, 3–14. Oxforad: Oxford University Press.

Weisberg M (2013) *Simulation and Similarity: Using Models to Understand the World*. New York: Oxford University Press.

Wikipedia contributors (2016) Blind men and an elephant. *Wikipedia, The Free Encyclopedia*, <https://en.wikipedia.org/w/index.php?title=Blind_men_and_an_elephant&oldid=752464208> (last revision: December 1, 2016).

Winsberg E (2010) *Science in the Age of Computer Simulation*. Chicago: University of Chicago Press.

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

Woody A (2015) Re-orienting discussions of scientific explanation: a functional perspective. *Studies in History and Philosophy of Science* 52: 79–87.

BALKAN JOURNAL OF PHILOSOPHY. Vol. 9, Issue 1, 2017

Invited paper

IS UNDERSTANDING FACTIVE? UNIFICATIONISM AND THE HISTORY OF SCIENCE

Sorin Bangu
Univ. of Bergen, Norway
e-mail: sorin.bangu@uib.no

Abstract:

Factivism is the view that understanding why a natural phenomenon takes place must rest exclusively on (approximate) truths. One of the arguments for non-factivism—the opposite view, that falsehoods can play principal roles in producing understanding—relies on our inclination to say that past, false, now superseded but still important scientific theories (such as Newtonian mechanics) do provide understanding. In this paper, my aim is to articulate what I take to be an interesting point that has yet to be discussed: the natural way in which non-factivism fits within the unificationist account of scientific explanation. I contend that unificationism gives non-factivists a better framework to uphold their position. After I show why this is so, toward the end of the paper I will express doubts with regard to the viability of de Regt's (2015) kind of non-factivism, based on the idea that understanding should be captured in terms of (scientific) skill.

Key words: understanding, factivity, unification, explanation.

1. Introduction

I take *factivism* about scientific understanding to be the view that understanding why a natural phenomenon takes place must rest exclusively on (approximate) truths; "quasi-factivism," a version of this doctrine, tolerates some falsehoods, but requires that they not have a predominant role in answering the why-question (Kvanvig (2003), (2009)). The main arguments for non-factivism—the opposite view, that falsehoods *can* play principal roles in producing understanding (Elgin (2009a, b))—fall roughly into two categories. They invoke either the extensive employment of idealizations in science, or our inclination to say that past, false, now superseded but still important scientific theories (such as Newtonian mechanics) do provide understanding.[1]

In this paper, I set the argument from idealizations aside, and will only deal with the second argument, from history. My aim is to articulate what I take to be an interesting point that (as far as I am aware) has yet to be discussed: the natural way in which non-factivism fits within the unificationist account of scientific explanation. I thus retain here sympathy for both unificationism (see my (2017)) and non-factivism (see my (2015)), although I will be critical about a recent version of the historical argument for this latter position,[2] due to Henk de Regt (2015). I believe that unificationism gives non-factivists a better motivation to uphold their position than to proclaim that it would be "unacceptable" and "absurd" (de Regt 2015, 3782; 3790) to deny that a surpassed, false theory provides understanding. More concretely, toward the end of the paper I will express doubts with regard to the viability of de Regt's own kind of non-factivism, based on the idea that understanding should be captured in terms of (scientific) skill.

My exploration of understanding here will proceed under several assumptions usually made (alt-

[1] Khalifa (forthcoming, esp. ch. 6) and Doyle *et al.* (forthcoming) give more precise formulations of the two views, and an extended bibliography on the factivism debate.

[2] In my (2017), I was also critical about the standard (Friedman / Kitcher) version of unificationism, and proposed a new, different "ontological" approach to it. I also argued that it is a refinement of Friedman's (1974) original idea, and not Kitcher's subsequent proposal, that accomplishes the task set for the unificationist project. Here, however, I will not rely on this new approach, but draw on the initial Friedmanian insight, and proceed as if it works (I will present an outline of it in section 3). I should add, although I cannot argue for this claim here, that I believe that the version of unificationism I advocate in my (2017) is perfectly compatible with everything I say in the present paper.

hough not universally agreed upon) in the philosophy of science literature, in particular in the debate on scientific explanation.[3] For one thing, I assume that the only way to gain understanding of a phenomenon p is by grasping an explanation of p[4]—and this is, I will postulate, the explanation accepted within the scientific community as the *best* available during the historical period when it was requested. Note that this is not to ask for a (the) *true* explanation, in the sense that all central assumptions in which understanding is grounded are true. The point of settling for less (as it were) is precisely to avoid prejudging the matter of factivity. Moreover, as the temporal indexing suggests, this requirement leads naturally to an examination of the historical record (the focus here will be on gravitational theories).

The structure of the paper is as follows. I shall begin, in section 2, with a discussion of the claim that "Newton had no understanding of gravitational phenomena" which de Regt (2015, 3782) finds "unacceptable." In section 3, I will briefly rehearse Friedman's (1974) unificationist account of explanation. In particular, I will insist that his distinction between "global" and "local" understanding (virtually forgotten today[5]) needs to be brought back into the debate. (I will also clarify why he believes that the only kind of explanation that yields understanding is of a unificationist type, and that understanding can only be of "global" kind.) However, my main objective in this section will be to argue for the non-factivist point that the intuition about the historical cases (e.g., that Newtonian mechanics does provide understanding) is best accommodated by the unificationist framework of explanation, while its causalist rival clashes with it. As we will see, this feature of unificationism is *not an accident*, since it turns out that the doctrine was designed with the challenge from history in mind. Finally, in section 4, I will take issue, as I said, with another non-factivist framework aimed to accommodate the historical cases, de Regt's (2015) "understanding as skill" proposal.

2. "Honorific" understanding? The curious case of Newtonian mechanics

In an insightful paper, de Regt (2015) discusses the argument from history, while also motivating his preferred version of non-factivism:

> In addition to the fact that scientists knowingly use unrealistic ('false') theories and models, the realist thesis[6] confronts a second problem: it appears to entail that past scientists who accepted theories we now have rejected as false, did not understand the phenomena at all. For example, Newton's theory of gravitation has been abandoned in favor of Einstein's theory of general relativity. Newton's theory is, from today's perspective, false and does not describe reality: the attractive forces that Newton postulated do not exist according to Einstein. But does this mean that Newton had no understanding of gravitational phenomena? That seems an unacceptable claim.
>
> So we face the dilemma of either giving up the idea that understanding requires truth or allowing for the possibility that in many if not all practical cases we do not have scientific understanding. I will argue that the first horn is preferable: the link between understanding and truth can be severed. This becomes a live option if we abandon the traditional view that scientific understanding is a special type of knowledge. While this view implies that understanding must be factive, I avoid this implication by identifying understanding with a skill rather than with knowledge. (2015, 3782)

It is not hard to see why a claim like "Newton did not understand gravitational phenomena" may raise eyebrows. Newtonian mechanics, and the classical gravitational theory in particular, are extensively taught in high-schools and universities, and form the backbone of numerous technological applications, as they get an enormous amount of predictions right (not only of the qualitative kind, but also in terms of precise numerical values). The Newtonian framework is almost universally recognized as a major step forward compared to Aristotelian teleol-

[3] A somewhat parallel debate takes place in epistemology, on the factivity of *knowledge* (roughly, the thesis that one cannot know p unless p is true). For some (e.g., Lipton (2001, 50)) understanding is a kind of knowledge ("of causes") and, if knowledge is factive (as generally accepted), understanding must be factive, too. Here I will put questions about the factivity of knowledge aside, but will get back to the relation between understanding and causation in section 3.

[4] Here I follow Strevens (2013) *contra*, for instance, Lipton (2009). As Strevens discusses, "grasping" is a difficult notion (I will not attempt to analyze it here), and may be rendered as "understanding that" (as in "understanding that the explanans leads to the explanandum"). Yet this is a sense of understanding different from the one of main interest here, used in the expression "understanding why p happens."

[5] Kitcher (1981) and (1989) accepts this, but does not use these labels.

[6] This is in essence the factivist thesis, "that science provides understanding of the world only if its theories are (at least approximately) true descriptions of reality" (2015, 3781). I inserted this footnote in de Regt's text.

ogy (*pace* Feyerabend and likeminded authors), and is surely better than Descartes' "physics" of tubes and vortices. And yet, just like the Aristotelian or the Cartesian fundamental postulates, the Newtonian fundamentals are false. Several of the key-notions and concepts are just not the kind of things they were assumed to be. Space and time are not absolute, and in fact they do not even exist as such, being "doomed to fade away into mere shadows, and only a kind of union of the two [the four-dimensional spacetime manifold] will preserve an independent reality," as Minkovsky memorably put it[7] (when reformalizing Einstein's Special Relativity). Simultaneity is not absolute, but only relative to a frame of reference. (Hence, what someone describes as "A before B" someone else is equally right to describe as "B before A"—utter nonsense from a Newtonian perspective.) Moreover, instant communication between different points in space is impossible (Newton thought that gravitation worked like that), being upper-bounded by the speed of light. (I leave out the more controversial issue of time-travel, apparently allowed by General Relativity, but laughable, and even metaphysically objectionable, in Newtonian terms.)

The list can continue, and one central item on it is what de Regt mentioned above: there is no force of gravity that pulls masses toward one another—and, in particular, that "bends" light. (In General Relativity, the gravitational effects are reconceptualized geometrically, as the influence of mass-energy on the geodesics). The light case is instructive, since—unlike many other cases in which the Newtonian numerical prediction is accurate—here it is significantly off the mark. Although the theory makes the true qualitative prediction that light will be deflected toward a massive object (since it assumes that light is composed of small material "corpuscles," subject to gravitation), the predicted angle of deflection is about *half* of the true, observed angle predicted by General Relativity.[8] So, in cases like these, given the extent to which both the central theoretical assumptions and the numerical predictions depart from what we take to be the truth, the doubts about the epistemic virtues of the Newtonian theory seem "acceptable": can we still say that it provides understanding of this individual phenomenon?

Despite all this, the factivist claim is still perceived as odd because it is felt to do an injustice to the Newtonian theory. One major undesirable consequence of calling the theory "false" is that in doing so we seem to treat it in the same way we treat the previous attempts meant to explain the same phenomena. On the other hand, we all realize that the Newtonian theory does not deserve this treatment; it is not "yet another false theory." Although it is false, it is so *in a different way* than its predecessors.[9] However, the question as to what this way is, specifically, remains open. In addition to making considerably more correct predictions than its predecessors, another reason may be its heuristic power, its role in the process of theory-invention. It seems impossible to imagine how Einstein, or anyone else, could have developed General (and Special) Relativity without reflecting on Newtonian ideas.

Factivists about understanding (like Kvanvig) feel the pressure coming from these cases, and agree that the notion can be applied after all. Nevertheless, they do so half-heartedly, since they also say that in these cases we talk about understanding in an "honorific" sense—and this is surely not the kind of understanding (and concession) that non-factivists are content with. A non-factivist, I take it, would want to claim that the Newtonian theory provides *genuine* understanding, if it does, *in the same sense* as General Relativity. Strevens, who otherwise displays non-factivist inclinations,[10] sides with the factivists here. His way out is by way of a semantical bifurcation as well. When it comes to such major, but superseded theories, he proposes we say that we (only) *understand with* them; he, too, introduces another sense of understanding, which, again, cannot be welcome by the non-factivists.

Although in the rest of the paper I will argue in favor of non-factivism, let me close this section with the suggestion that the non-factivist attitude of almost absolute deference to Newtonianism (such as de Regt's) seems to me exaggerated. I suspect that this attitude may in part be due to historical myopia. It is then helpful to recall that the doubt as to whether this theory offers understanding is actually very

[7] Cited in Jammer (2006, 300).

[8] In Query 1, at the end of his *Opticks* (1704), Newton speculated that light particles, even if of small mass, are affected by gravity just as ordinary matter. In 1804, von Soldner published the first calculation of a deflection angle (by the Sun)—about 0.9 seconds of arc. Another, better-known example of predictive disagreement is the advance of Mercury's perihelion, but there the discrepancy is not as big as in the light case. In 1859, Le Verrier calculated the rate of precession predicted by the Newtonian theory, and noted that the observed rate (ca. 570 arcsec/century) is larger by ca. 43 arcsec/century, a discrepancy that was corrected when the prediction was made using General Relativity.

[9] De Regt (2015) correctly alludes to this idea throughout his paper, but the solution he offers I believe to be flawed (see section 4).

[10] Strevens (forthcoming) sets out to explicate how idealizations produce understanding.

old. As is well known to historians of science, the consensus on the explanatory virtues of Newton's theory took a while to build, despite his contemporaries' full awareness that it yielded precise predictions. Interestingly, their perspective on the theory, and their reasons to doubt its epistemic virtues were different from ours – they were, we would say today, much more metaphysically prejudiced. Our perspective, I take it, is something like the following: "were the theory true, it would explain and provide understanding; but it is not, and yet, given all we know about how theories in physics come and go, it is legitimate to ask if it can still give us understanding." That is, we have no problem in counting the theory among the contenders—it is just that in the end the world ceased to cooperate. However, those who doubted the theory in the 17th century (Huygens among them[11]) took a much harsher stance: they seemed to be saying that the theory could not even have been a contender among the theories able to provide explanations and understanding! This was so since, they asserted, Newton presented a mere "mechanics" (and not a "physics"!)[12] by "simply posit[ing] mathematical forces, which pertains to a different order of things, namely geometry, and thus could not constitute a true explanation of physical phenomena" (Gingras 2001, 399).[13]

What I suggest, then, is that we should be a bit more relaxed when looking into this matter. It is not "absurd" to doubt the virtues of Newtonianism; our intellectual ancestors did it, and we are no less entitled to this than they were. But, first, to take this doubt seriously is not to say that the theory is worthless. Second, it is not to say that, once demoted to a falsehood, the theory is on a par with a number of other far-fetched speculations that are also nominally false. This is so, I submit, despite the fact that, as we saw, it is an unresolved issue how to spell out these two intuitively correct claims. Hence the suggestion: it may be that we could not make progress on this issue because, as the unificationists urged, we have been looking at the relation between explanation and understanding in the wrong way all along.

3. Unificationism, Causalism, and Non-Factivism

I am now going to try to make a case for non-factivism within a unificationist framework. That is, I will argue that it is possible to say that the Newtonian gravitational theory (and the Newtonian framework in general) provides understanding—despite being strictly speaking false. In addition (and importantly), the argument will not involve any semantical bifurcation: we will be able to say that the Newtonian gravitational theory provides understanding *in the same sense* in which we say that General Relativity provides it. As I noted at the outset, this possibility opens up once we adopt unificationism about explanation. Therefore, at this point we need to take a detour and give a brief exposition of this account; afterwards, I will elaborate on how the accomplishments I just listed are achieved.

However, before we get into this, let us note that the possibility I mentioned above did not open up by accident. On the contrary, Friedman built it into unificationism right from the beginning, when he first proposed the idea in his 1974 master paper. He set the aim of unificationism as follows, explicitly putting the "the theories of Newton and Einstein" on the same footing:

> [I]t seems to me that it would be desirable, if at all possible, to isolate a common, objective sense of explanation which remains constant throughout the history of science; a sense of 'scientific understanding' on which the theories of Newton, Maxwell, Einstein, and Bohr all produce scientific understanding.
>
> [An account of explanation] should be sufficiently general – most, if not all, scientific theories that we all consider to be explanatory should come out as such according to our theory. Although it is unreasonable to demand that a philosophical account of explanation should show that every theory that has ever

[11] It seems that Huygens' objection was that gravitational attraction cannot be an "inherent property" of matter. See Mischel (1966, 40).

[12] The contrast is in the review of *Principia* in *Journal des Savants* in 1688. According to Gingras (2001, 399), the review "praised the author's geometry but concluded that to make his work 'as perfect as it was possible' Newton now had to produce 'a physics as exact as his mechanics', which, it was added, could be done only by substituting 'true motion in place of those he has supposed.'"

[13] Things are even more complicated, as Mischel (1966, 42) points out that "Newton's speculated on the possibility of explaining gravity in hydrostatic terms as due to variations in the density of the 'aethereal medium' … , but was dissatisfied with this explanation because it seemed to lead to absurd consequences … . He also speculated on the possibility that the small particles of bodies have 'powers, virtues, or forces, by which they act at a distance' … , but elsewhere holds that it is 'inconceivable' that matter should 'operate upon and affect other matter without mutual contact', rejecting the notion that gravity is 'innate, inherent, and essential to matter' as an 'absurdity.' (Letter to Bentley, quoted in Cajori's Appendix to the *Principia*, vol. HI, p. 634). While Newton's statements are guarded so that his views on this question are hard to determine, most recent commentators agree that he cannot be regarded as a 'Newtonian.'"

> been thought to be explanatory really is explanatory, it must at least square with most of the important, central cases. (1974, 13)

This objective sense is meant to be captured in the unificationist framework, as we will briefly see below.[14] To make a start, let us think of explanation in the most general terms, as the usual triple <explanans | dependence relation (*d*) | explanandum>. The explanandum is a phenomenon (*p*) to explain, the explanans is the set of assumptions based on which we explain it, and relation *d* makes explicit how *p* depends on the explanans.[15] Less formally, such a dependence can be cashed out as logical (à la Hempel), causal or statistical (à la Salmon), or in other ways. For instance, in Hempel's well-known deductive-nomological account, to explain a phenomenon is to derive it as the conclusion of a sound deductive argument from one or more laws of nature, with the help of some initial conditions. (The explanans is thus a composite, with all its components required to be true—so there is no doubt that Hempel was a factivist! Yet, as is well known, the account fell through because of several counterexamples, e.g., the flagpole/shadow.)

Now, if (as assumed) understanding of a phenomenon comes (only) from an explanation of it, and explanations have the structure <explanans | *d* | *p* (explanandum)> how does understanding arise? How is unification playing any role here? Let us take schema E below to capture (again, in the most general and non-committal terms) what it is to explain a phenomenon:

> Phenomenon p is explained iff both:
> i) the explanans is identified
> ii) relation *d* is specified

Friedman formulates the problem of understanding as the question Q_E below:

> How does a scientific explanation (of phenomenon *p*) produce *objective understanding* (of *p*)?

Schema E suggests a way to address this question: supplement these two conditions with additional ones. For instance, one may require adding clauses to the effect that the explanans has certain features[16]—and, as we will see, there are several natural candidates to consider here. One of them involves the notion of unification, and will arguably do a better job than the others. The notion of unification I will be working with is Friedman's,[17] and will be the third option.

In looking for additional requirements, the following thought (a kind of "closure" principle) is very tempting: understanding p may come from the fact that we already understood the explanans (so when *d* is specified, it is required that it be able to "transmit" this understanding). Thus, one idea is to replace i) with the following clause in schema E:

> i') the explanans is familiar

(Some authors require that the explanans be "less in need of an explanation" than *p*.)[18] If the explanans yield *p*, and the former is familiar to us, then we seem entitled to claim that we understood the latter. Yet this "reduction" will not do, not only because of immediate counterexamples (sugar's solubility in water—a familiar sight—is explained by the unfamiliar properties of the interaction of water and sugar molecules), but also because familiarity is fundamentally subjective, far from the objective notion Friedman seeks.[19]

Another appealing idea, this time not relying on any transitivity, is to bring in the notion of cause. So, we replace i) with:

> i'') the explanans is the cause of *p*

Now, intuitively, to understand why *p* happens is to identify what causes it. So, for a phenomenon

[14] The next several paragraphs draw on material from my (2017), in which I articulate and defend a *version* of unificationism which departs in several respects from the standard doctrine I am presenting here. See footnote 2.

[15] Specifying the nature of relation *d* is also important in order to rule out an explanation of a phenomenon *p* by "deriving" it from God's omnipotence. This is what many religious people invoke, yet they are notoriously unable to give details as to *how* God makes *p* happen (hence one sense of the proverbial saying that God works in mysterious ways).

[16] As I discuss in my (2017), one may require, as a different strategy, that the instantiations of *d* satisfy certain constraints. Kitcher (1981), (1989) implements this strategy by making appeal to a different notion of unification.

[17] I leave out Kitcher's strategy for reasons I explain in my (2017). In essence, I argue that Kitcher is not interested in how understanding is generated from an explanation; in this matter, he defers to the account offered by Friedman (although he criticizes it for formal mistakes).

[18] As Schurz (1999, 97) put it; see also Schurz and Lambert (1994, 105).

[19] Another (but related) idea Friedman discusses is borrowed from Toulmin (1961), and requires that the explanans be an "ideal of intelligibility" adopted by the scientific community at a certain historical moment. Hence, we should add this clause: the explanans is an ideal of intelligibility. Does this addition answer Q_E? Not really. Although, just as above, one may feel she understood the explanans—because, by hypothesis, everybody around her feels the same way—this is ultimately a matter of psychological (or sociological) fact, and not objective, as Friedman wanted.

we seek to understand we need to find *what* "produced" it, as well as to demonstrate some details about *how* it came about, by identifying the relation *d* in the form of a causal chain.

This idea is admittedly immensely popular.[20] Under the usual assumption that causal relations are objective, it would have surely answered Q_E if it were not for the formidable empiricist-inspired arguments against this notion. Briefly put, the problem the critics see is that we do not quite know what (the second part of) condition i'') even means, or what this clause gives us in addition to a mere word ("cause.") In a Humean spirit, Russell became the founder of what is today called *causal eliminativism*,[21] by famously assigning causation the undignified status of a "relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm" (1912, 1). Later on, Quine also embraced eliminativism.[22] Recent authors are still uncomfortable with this notion; Arif Ahmed, for instance, believes that "Causation is a pointless superstition" (2014, vii).

Nevertheless, given the naturalness of condition i''), it is hard to dismiss it—and in my (2017) I do not dismiss it, despite ample evidence from (the history of) philosophy and science that it was never spelled out. Moreover, I take it that at present there is no consensus on what "A causes B" means, despite the efforts of many important philosophers. Yet I do not find it surprising that the debate on explanation centered on it. Between roughly the 1950s and 1970s, the philosophers of explanation who resisted Hempel's influence (for him, understanding was ultimately a subjective-psychological business) had no other option than to first advance a theory of causation. That is, they wanted to first spell out i''), and then, on this basis, work out a theory of explanation, which could not but be a *causal* theory of explanation. This was exactly Salmon's project in the 1970s and 1980s, part of a more general revival of interest in metaphysics led by Kripke and Lewis. Yet others—Friedman and Kitcher prominent among them—realized that a different strategy was also viable: account for scientific explanation *without* first toiling on a theory of causation, i.e., by avoiding sinking (yet again) into the swamps of metaphysics. Thus, unlike Salmon, they were still following the logical empiricists in their anti-metaphysical orientation; but this theory of explanation, as the Friedman quote above reveals, was meant to account for *objective* understanding (this time parting ways with Hempel). I emphasized this dialectic in my (2017), and I retain the observation here, since it is not always explicitly discerned in the literature. I submit that this captures the spirit of unificationism: a salutary attempt to separate the task of finding a theory of scientific explanation from the uncertain fortunes of the task of elaborating a theory of causation. Next, I will present some details of the unificationist proposal, and I will only consider Friedman's[23] (somewhat against the received view[24]).

So, given the background I have provided so far, I read Friedman's paper as first proposing a specific take on the problem of understanding, and second, as formulating an answer to it. I will begin with the first point.

More precisely, I read Friedman as saying that if Q_F above is a formulation of the problem of scientific understanding, then the problem is insoluble—since, as we just saw, none of the natural, immediate proposals above really works. But, he points out, we also have the option to drop a certain assumption underlying Q_F—namely, that when explaining phenomena we tackle them one by one, in an *individual*, piecemeal fashion (this would be, following Friedman, the assumption of "locality"). Now, if we drop this assumption, we say that science explains in a "wholesale" fashion, and thus "understanding" means "global understanding," or understanding *the world*, as a whole. This sounds grandiose, but the idea of *global understanding* is actually mundane and natural. In essence, achieving this kind of understanding comes as a result of decreasing the (total) number of why-questions on the scientific agenda. It is reasonable to assume that a world posing fewer why-questions, i.e., containing fewer "mysteries," is *more comprehensible*.

Thus, under Friedman's reformulation, the relevant question reads as follows (I will label it Q_F):

> How does a scientific explanation generate *objective global understanding*?

As different from virtually any other rival account, unificationism takes Q_F to be the main con-

---

[20] Giving a list of its advocates would be too long; for a recent take on it, see Grimm (2014). I talk loosely about 'the' cause, although of course causes may not be unique.

[21] According to Shaffer (2016), "The main argument for eliminativism is that science has no need of causation. The notion of causation is seen as a scientifically retrograde relic of Stone Age metaphysics."

[22] He wrote: "[T]he notion of cause itself has no firm place in science. The disappearance of causal terminology from the jargon of one branch of science and another has seemed to mark the progress in the understanding of the branches concerned" (1976, 242).

[23] I have dealt with Kitcher's in my (2017). See Schurz and Lambert (1994) for a carefully crafted alternative to Friedman and Kitcher.

[24] See Woodward's 2011 *Stanford Encyclopedia of Philosophy* article, where Friedman's position is barely mentioned.

cern in the philosophy of explanation; unlike Q_E, it is answerable. The unificationist deals with it by imposing a different requirement on the explanans, that it must be "*more comprehensive*" (Friedman 1974, 19; italics in original). The requirement amounts to asking that the explanans serve as an explanatory basis not only for the derivation of the phenomenon we sought to understand initially, but for (many) *other* phenomena as well.[25] Such a comprehensive explanans will be a *unifier* of these various phenomena. Thus, the unificationist proposal to complete schema E is to replace i) with:

i''') the explanans is a unifier

By doing this, we have now achieved the ability to answer Q_F, i.e., to show how we get global understanding; we can also show that it satisfies the objectivity criterion. In order to clarify why this is so, let us begin by making the additional (and trivially true) assumption that the scientific agenda contains a multitude of puzzling phenomena; call them p_1, p_2, p_3,… Now, diagram (A) below represents the received view on explanation: different phenomena are explained, and thus made intelligible *individually*, by different explanans. Yet *if* we happen to be in a situation in which the explanans of one phenomenon (say p_1) is comprehensive, or a unifier—we will call it U_{exs}—that means that in addition to its explanation $\langle U_{exs} \mid d \mid p_1 \rangle$, other explanations, of the form $\langle U_{exs} \mid d \mid p_2 \rangle$, $\langle U_{exs} \mid d \mid p_3 \rangle$, etc. can be formulated as well. This felicitous situation is captured in diagram (B). The unificationist contends that the epistemic situation described by diagram (B) is epistemically superior to the situation described by diagram (A):

(A)	(B)
$\langle \text{explanans}_1 \mid d \mid p_1 \rangle$ $\langle \text{explanans}_2 \mid d \mid p_2 \rangle$ …	$\langle U_{exs} \mid d \mid p_1, p_2, \ldots \rangle$

Diagram (B) describes *unificatory explanation*, i.e., explanation-via-a-unifier. As is hopefully clear from diagram (B), when explaining p_1, p_2, p_3, etc., we do not need several separate explanans (i.e., corresponding to each phenomenon), but only one (U_{exs}). Hence, in the end, only one unexplained item ("mystery") remains, instead of several,[26] and we did decrease the number of phenomena we must accept as unexplained ("mysterious"). An important consequence relevant for the previous discussion is that failures to account for some *individual* phenomenon do not really matter: we now think of explanation and understanding "wholesale," not "retail." Thus, if we are ever in the situation described by diagram (B), we live in a world containing fewer mysteries; hence, says Friedman, such a world is *more* comprehensible, and we gained global understanding.

In addition to noting that (global) understanding comes out now as a *gradual* notion, we should also remark that, since we are after achieving a *quantitative* reduction in the number of why-questions generator phenomena, we can address the objectivity issue. Anyone who can count[27] can compare the situations in diagrams (A) and (B): in (A) we are left with two questions ("why does the explanans of p_1 (p_2) hold?"), while in (B) we have only one question left to answer ("why does U_{exs} hold?"). Thus, no matter what one happens to be familiar with, or what kind of metaphysical preferences (or prejudices) one may entertain during a certain historical period (recall the issues with Newtonianism above), one must accept that this quantitative, gradual increase in global understanding is an objective fact. A world with few(er) mysteries is objectively more comprehensible than one with many.[28] So, to summarize, for unificationists,

(a) explanation is explanation-via-a-unifier, and provides
(b) objective global understanding.[29]

[25] Standard unificationists (Friedman and Kitcher) follow Hempel in requiring that *d* be a logico-mathematical derivation, and here I will accept this. However, the version of unificationism I adopt in my (2017) leaves the nature of *d* open.

[26] Once a unifier U_{exs} is found, scientists will presumably treat it as an object of further investigation (a "mystery") as well, and will try to integrate it in an even more encompassing system.

[27] This is an important juncture in the debate on the viability of unificationism, as the question "how do we count (brute) phenomena?" (first raised by Salmon) is taken to create serious difficulties for Friedman (and gave Kitcher the motivation to propose the version of unificationism in which the reduction is in the number of explanatory *patterns*, not phenomena).

[28] This is meant to address one of Morrison's (2000, 26-29) criticism of Friedman's account, that it is not objective.

[29] From this point onward, I will be writing as if I accept this standard, traditional version of unificationism, yet let me stress (and elaborate on) a point I made in footnote 2: this standard (Friedman/Kitcher) unificationism is *not* the version of the doctrine I ultimately accept. In my (2017), I agreed that the standard version has fatal problems, and I even spelled out a series of novel objections to it; the essential one regarded point (a) here, that for the standard view understanding is *only* global, and thus no room is left for a local sense of understanding. I also offered a reformulation of the position and tried to show that this refurbished version can incorporate a local sense of un-

We now have all the elements to make the case for the claim above, that the doctrine provides (in Friedman's words) "a sense of 'scientific understanding' on which the theories of Newton [and] Einstein … produce scientific understanding."

Why is this so?—again, despite the fact that Newton's gravitational theory is false. The answer should now be clear: although false, the theory qualifies as explanatory insofar as it is unificatory. And since there is no doubt that it is so, it follows that it provides understanding—where, again, this means, specifically, global (not local/individual) understanding, the (only) sense of understanding that this standard unificationism recognizes. Moreover, the very same claims can be made about Einstein's theory—and note that these claims will hold even if, at some point in the future, his theory is superseded. Its fundamental postulates also unify, to an even larger extent than its Newtonian ancestor. General Relativity applies not only to planets, comets, and stars,[30] but also to much more, including "exotic" phenomena (about which the Newtonian theory has basically nothing to say, or cannot even recognize)—such as the extreme cases of the "bending" of light (i.e., "black holes") or the expansion of the Universe.

In short, the unificationist is in a position to maintain that since both theories are major unifiers, they are both explanatory and hence providers of understanding in the same global sense. Furthermore, one may also apply this idea to other past, superseded theories, such as the phlogiston theory of combustion, or the caloric theory of heat. Are they explanatory too? Did the advocates of the phlogiston theory—beautiful minds such as Stahl and Priestley—have any scientific understanding at all? de Regt finds giving a negative answer (only) "awkward" (2015, 3792)—and I agree (yet for different reasons; see section 4). A unificationist also answers in the affirmative, as follows: their theory did yield understanding *proportional to its unificatory power*, power which surely cannot be denied. (On the other hand, and what mattered most, this power was surely less impressive than that of the rival oxygen theory.)

We thus see that unificationism deals with such difficult historical cases naturally, without having to engage in the semantical contortionism of the factivists. For a unificationist, there is no need to talk about "honorific understanding," or "understanding with"[31] when confronted with (odd-sounding) pronouncements such as "Newton did not understand gravitational phenomena." Thus, unificationism squares quite nicely with non-factivism, and does so elegantly, by way of principled argument, not by mere foot-stomping (or by labeling other stances as "absurd" "unacceptable," etc.)[32]

To close this section, it is instructive to compare unificationism and causalism along the factivist dimension. We saw that unificationism is non-factivist by design, and hence it naturally resists the bothersome claims denying most of the greatest scientists any understanding of the world. (Einstein will be denied this too, once his theory is superseded!). The causalist, on the other hand, is in a more uncomfortable position. Recall that the causalist contends that to explain, and to understand, a phenomenon is to identify its cause—and now we can also add, she is claiming this in a "localist," individualistic manner. If so, since Newton was surely wrong to single out the force of gravity as the cause of the planets' elliptical orbits around the Sun (to consider his most famous achievement), then the causalist will have to maintain that he did not precisely understand the phenomenon for whose understanding we celebrate him—and thus she is back in the situation of accepting the "unacceptable."

4. De Regt's skill-based non-factivism

Unlike causalism, de Regt's "skill" account mentioned at the outset goes against "the realist [i.e., factivist] thesis that science can provide understanding of the world only if its theories and models are (at least approximately) true descriptions of reality," by invoking "the practice and history of science," which "reveal that understanding can be—and is in fact often—achieved through models that are unrealistic, highly idealized representations of the target system, or on the basis of theories that are strictly speaking false, for example because they have been superseded by theories with radically different ontological assumptions." He rejects that idea that "the traditional view of scientific understanding should be replaced by an alternative interpretation that allows for understanding without truth," and offers

derstanding, while also being able to deal with the typical objections (the conjunction problem, etc.). Moreover, I argued that the new version is in fact compatible with causalism (if the doctrine is to ever be articulated properly), and that it takes precedence over it.

[30] Salmon's balloon explanation (1990, 11) is a nice example of an application of the theory to terrestrial phenomena.

[31] The split between local and global understanding may sound like a manoeuvre of semantical nature as well. Yet, the unificationist points out, the introduction of this distinction was principled, and not an ad hoc decision when facing recalcitrant cases.

[32] To insist on a point I made above, I believe that the version of unificationism I advocate in my (2017) can accommodate this non-factivist orientation—although unfortunately I do not have the space to argue for this here.

> …an alternative by characterizing understanding as a skill rather than as a kind of knowledge. On my account, scientific understanding does not require that theories are (approximately) true but instead that they are intelligible, that is, they should have qualities that allow scientists to *use* them for constructing models of the phenomena. Accordingly, intelligibility is a pragmatic value that depends on the context: theories are employed by scientists with specific skills, for particular explanatory purposes, and which theories are useful depends on these contextual factors. (2015, 3794-5)

The central idea de Regt puts forward is that a theory yields understanding (of a phenomenon) when the theory (under which the phenomenon falls) is "intelligible." There is rather little that one can prima facie say when assessing this point. *Of course* theories must be intelligible; after all, how would an unintelligible theory find a place in science? Moreover, what would an unintelligible theory look like? I am not aware of any concrete example of such a theory, and I do not think de Regt provides one, although it would have helped, just to grasp the contrast better. (True, scientists say sometimes, disparagingly, that a rival's theory is "unintelligible," but I do not think that this anecdotal sense is intended here.)

Thus, intelligibility is the key-notion of the account. What is it, then? It occurs when scientists possessing "specific skills" can *use* the theory for "particular purposes," and in various contexts, to construct "models of phenomena" (where this presumably means predictively correct models). Again, the account seems overgenerous: how could a theory that *cannot* be used in these very general ways ever be taken seriously, or even deserve the name of a "theory"? Equally troublesome is the converse thought: on these lax terms, *any* kind of "story" about the world, no matter how arbitrary and ad hoc, may come out as yielding understanding. If the "contextual factors" are decisive, even a religious doctrine (any of the known ones) yields understanding in the hands of a bunch of clever priests, since they can make the world "intelligible" to their credulous followers.

However, a bit earlier in the paper there is more about "intelligibility" (see also de Regt (2009)):

> Explanatory understanding of phenomena requires that theories are intelligible, where intelligibility is defined as the value that scientists attribute to the cluster of qualities of a theory that facilitate its use. It follows that intelligibility is not an intrinsic property of a theory but an extrinsic, relational property, because it depends not only on the features of the theory but also on the skills of the scientists who work with it. Theories are not intrinsically intelligible or unintelligible, but intelligible or unintelligible for a particular scientist or group of scientists. In other words, intelligibility is a context-dependent value related to the skills of scientists. (2015, 3793)

To recap: a theory generates understanding to the extent that it is intelligible, and intelligibility is in turn determined by whether or not particular groups of scientists (possessing certain skills) are able to use the theory to construct empirically adequate "models of phenomena." On one hand, this is fine, as it speaks against factivism: if the scientists are able to do this (and oftentimes they are indeed), then the truth of the theory does not really matter. On the other, however, I think the above paragraphs also entail that it is just fine to say that a theory loses its capacity to yield understanding if some other groups of scientists (presumably lacking the first group's skills, while perhaps having others) are not able to use it in this way. If the intelligibility of a theory is a "context-dependent value," related to (someone's) skill, then so is its capacity to explain and produce understanding.

This is worrisome; strong intuitions push in the opposite direction. A hammer is no less a hammer because a small child is too feeble to lift it up and hit a nail; moreover, it is not the case that if a shoe is used to drive a nail, then it suddenly becomes a hammer. What is worrisome here, then, is the almost extreme subjectivism and relativism of this contextualist account of understanding—in stark contrast, I should add, not only to unificationism, but also to causalism. Such a contextualism seems compatible with the strange situation in which different groups of scientists work with the same theory and make opposite claims with regard to its capacity to provide understanding of the same phenomenon—because they possess different skills. By way of comparison, this is immediately ruled out by the other two views on explanation. Take causalism: the cause of p is still its cause no matter what group of scientists one asks, and what skills they have. Similarly, whether or not a theoretical framework unifies a series of phenomena is an objective fact, too. Of course, in both cases some (groups of) scientists may not *realize* that the respective facts hold—depending, presumably, on what skills they have—but their skills should not become determinant in how the theory is characterized (as capable of yielding understanding, or not). It is rather puzzling to contend that explanation and understanding are not so much about theories, but *theorists*. I am afraid that the dialectical move here, defending non-factivism by embracing a relativist-subjectivist contextualism, is not viable.

Acknowledgments. I thank the editor, Lilia Gurova, for the invitation to write for this special issue, and Kareem Khalifa for extended conversations on these topics. All responsibility for the final version is mine.

References

Ahmed A (2014) *Evidence, Decision and Causality*. Cambridge: Cambridge Univ. Press.

Bangu S (2017). Scientific Explanation and Understanding: Unificationism Reconsidered. *European Journal for Philosophy of Science* 7(1): 103–126

Bangu S (2015). Why does Water Boil? Fictions in Scientific Explanation. In *Recent Developments in the Philosophy of Science* eds U. Mäki et al., 319–330. Springer.

de Regt, H (2009). Understanding and scientific explanation. In *Scientific Understanding. Philosophical Perspectives,* eds de Regt H., S. Leonelli, K. Eigner, 21–42. Univ. of Pittsburgh Press.

de Regt H (2015). Scientific understanding: truth or dare? *Synthese,*192: 3781–3797

Doyle Y, Egan S, Graham N, Khalifa K (forthcoming). Non-Factive Understanding: A Statement and Defense. *Journal of General Philosophy of Science* (Special issue on Naturalism, Normativity, and Understanding)

Elgin C. (2009a). Exemplification, idealization, and scientific understanding. In Suárez (2009). 77–91.

Elgin C (2009b). Is understanding factive? In *Epistemic value, eds* Haddock A., A. Millar, D. Pritchard, 322–330. Oxford Univ. Press.

Friedman M (1974). Explanation and scientific understanding. *Journal of Philosophy* 71: 5–19.

Gingras Y (2001) What did Mathematics do to Physics? *History of Science* 39: 383–416

Grimm S (2014). Understanding as knowledge of causes. In *Virtue Scientia: Essays in philosophy of science and virtue epistemology,* ed A. Fairweather. Special issue of *Synthese*.

Jammer M (2006). *Concepts of Simultaneity: From Antiquity to Einstein and Beyond*. Johns Hopkins Univ. Press.

Khalifa K (forthcoming). *Understanding, Explanation, and Scientific Knowledge*. Cambridge Univ. Press.

Kitcher P (1981). Explanatory unification. *Philosophy of Science*, 48: 507–31.

Kitcher P (1989). Explanatory unification and the causal structure of the world. In *Scientific Explanation,* eds P. Kitcher, W. Salmon, 410–505. Univ. of Minnesota Press.

Kvanvig J. (2003) *The value of knowledge and the pursuit of understanding*. Cambridge Univ. Press.

Kvanvig J (2009) Responses to critics. In *Epistemic value,* eds A. Haddock, A. Millar, D. Pritchard, 339–352. Oxford Univ. Press.

Lipton P (2001) What good is an explanation? In *Explanation: Theoretical Approaches and Applications,* eds Hon G., Rakover S., 43–61. Synthese Library, Springer.

Lipton P (2009) Understanding without Explanation. In *Scientific Understanding: Philosophical Perspectives*, eds de Regt H., Leonelli S., Eigner K., 43–63.Pittsburgh Univ. Press.

Mischel T (1966) Pragmatic Aspects of Explanation. *Philosophy of Science* 33(1): 40–60.

Morrison M (2000) *Unifying Scientific Theories. Physical concepts and mathematical structures*. Cambridge Univ. Press.

Quine W V (1976) *The Ways of Paradox and Other Essays* (Revised and enlarged ed.) Harvard Univ. Press.

Russell B (1912) On the notion of cause. *Proceedings of the Aristotelian Society,* 13: 1–26. New Series.

Salmon W (1990) Scientific explanation: Causation and unification. *Critica* 22(66): 3–21.

Schaffer J (2016) The Metaphysics of Causation. *The Stanford Encyclopedia of Philosophy* (Fall 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics/>.

Schurz G (1999) Explanation as unification. *Synthese*, 120: 95–114.

Schurz G, Lambert K (1994) Outline of a theory of scientific understanding. *Synthese* 101: 65–120.

Strevens M (2013) No understanding without explanation. *Studies in History and Philosophy of Science*, 44: 510–555.

Strevens M (forthcoming). How Idealizations Provide Understanding. In *Explaining Understanding,* eds S. R. Grimm, C. Baumberger, S. Ammon. Routledge.

Toulmin S (1961) *Foresight and understanding*. London: Hutchinson.

Woodward J (2011). Scientific explanation. In *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta, (Winter 2011 Edition). URL = http://plato.stanford.edu/archives/win2011/entries/scientific-explanation/.

ON SOME NON-TRIVIAL IMPLICATIONS OF THE VIEW THAT GOOD EXPLANATIONS INCREASE OUR UNDERSTANDING OF EXPLAINED PHENOMENA

Lilia Gurova
New Bulgarian University, Sofia, Bulgaria
e-mail: lgurova@nbu.bg

Abstract:

The central argument in this paper is the following: if we agree that one of the aims of explanation is to provide or increase understanding, and if we assess understanding on the basis of the inferences one can draw from the knowledge of the phenomenon which is understood, then the value of an explanation, i.e. its capacity to provide or increase understanding of the explained phenomenon, should be assessed on the basis of the extra-inferences which this explanation allows for. The extra-inferences which a given explanation allows for constitute its inferential content. The analysis of the explanation's inferential content could be applied to all kinds of explanations with the aim of assessing their goodness. I show how such an analysis helps us to better understand a number of difficulties that have puzzled contemporary philosophers of explanation: the flagpole counterexample to the deductive-nomological model of explanation, the conjunction problem, the difference between good and bad circular explanations.

Key words: explanation, understanding, inferential content, the conjunction problem, circular explanations.

1. Introduction

There is a well-recognized revival of interest in understanding in recent philosophy of science[1]. The beginning of this revival was marked by the publications of several prominent philosophers (von Wright, 1971; Toulmin, 1972; Friedman, 1974) who challenged the view that had dominated the field for more than a half of century, the view that understanding is an epistemically unimportant, either pragmatic, or subjective (psychological) side product of explanatory activities. In the last 15 years, much effort has been put into illuminating the epistemic significance of understanding (de Regt, 2009; Kvanvig, 2009) and its proper relation to explanation (de Regt, 2013; Khalifa, 2013; Gervais, 2015). However, an important asymmetry between explanation and understanding has received little attention so far. The asymmetry lies in the fact that when we want to explain P, we usually look for what determines P, or what makes P possible (Lipton, 2004). At the same time, when we try to find out whether one understands P, we usually look for how (s)he uses her/his knowledge of P for "going beyond the information given" (Bruner, 1957); in particular, we look for what kind of inferences (s)he can draw from the representation of P and the available P-relevant knowledge. Taken together with the assumption that good explanations enhance our understanding of explained phenomena (Friedman, 1974), awareness of this asymmetry

[1] (de Regt, Leonelli & Eigner, 2009) is representative of the state of the art in this new field in philosophy of science. See also (Khalifa, 2012; 2015), (Newman, 2014), (Bangu, 2016), and (Grimm, Baumberger & Ammon, 2017) for important follow-ups.

leads to the following simple argument: if one of the main aims of explanatory activities is to provide or increase understanding, and if we assess understanding on the basis of the inferences one can draw from the knowledge of the phenomenon which is understood, then the value of an explanation, i.e. its capacity to provide or increase understanding of the explained phenomenon, should be assessed on the basis of the extra-inferences which this explanation allows us to draw. This simple argument leads to a view on explanation which differs from the existing views in philosophy of science in an important aspect: instead of focusing on the structure of explanation—and in particular, on the connection between the *explanans* and the *explanandum*—this view invites us to take into account what kind of new inferences a given explanation allows for. The latter suggestion has several important implications, and the aim of this paper is to discuss them and show where their significance lies.

2. A new model of explanation, or a complement to the existing models?

The proposal to assess explanations on the basis of their inferential content, i.e. to assess them on the basis of the inferences they allow for, does not compete with the existing models or theories of explanation, i.e. with the different versions of causal, mechanistic, and unificationist models. This is because the proposal to assess the explanations' inferential content does not address the question "What is the proper explication of the explanatory relation (the relation between the *explanans* and the *explanandum*)?"—a central question for the currently popular models of explanation in philosophy of science[2]. In this sense, the proposal which is suggested here is compatible with any of the models that are explications of the *explanans-explanandum* structure. At the same time, the proposal to assess the explanations' inferential content is an important complement to the existing explicatory models of explanation, as it supplies them with a tool for distinguishing their good from their bad instantiations[3]. The proposal to assess the explanations' inferential content builds on the idea that the poor inferential content of a tentative explanation is a sign that the latter is a bad explanation, even if it satisfies all the requirements of a particular explicatory model of explanation. Thus the examination of the inferential content allows one to show "why the big bang is a poor explanation for cultural differences in color preferences" (Lombrozo, 2013, p. 271), even though the big bang event matches all the conditions for being recognized as a distal (ultimate) cause of everything that is real today (including cultural differences in color preferences). If we examine the inferential content of the big bang explanation, its deficiencies become clear immediately: from this explanation we cannot draw any interesting inferences about the cultural differences in color preferences, i.e. the part of the inferential content of the big bang explanation which is relevant to this particular *explanandum* (the cultural differences in color preferences) is very small. The latter fact tells us why referring to an ultimate cause as an explanation of cultural color preferences is so unattractive.

[2] The deductive-nomological (DN) model of explanation construes the relation between the *explanans* and the *explanandum* as a logical entailment. The various causal models of explanation, as their name suggests, claim that the *explanans* should refer to the cause of the explained event, phenomenon, etc., and that the *explanandum* should describe the explained event or phenomenon. The different versions of the unificationsit model of explanation could be viewed as generalizations of the DN model: the *explanandum* should be inferred from the *explanans*, but the inferential relation between them is not necessarily one of entailment.

[3] Each of the existing models of explanation has good as well as bad instantiations. The DN model, for example, cannot recognize the explanation of the height of a flagpole by the length of its shadow as a bad one (see section 3.). The causal models cannot rule out as bad explanations some of the explanations that refer to distal causes, e.g. the big bang explanation of the cultural differences of color preferences. For a recent discussion on some of the bad instantiations of unificationist models, see (Petkov, 2015).

3. The proposal to assess the explanations' inferential content helps to better understand some famous controversies in philosophy and science

The flagpole counterexample to the deductive-nomological model of explanation

The authorship of the flagpole counterexample to the DN model of explanation is usually attributed to Sylvain Bromberger (1966)[4]. Briefly, this counterexample states that although the height of a flagpole could be derived from the length of the flagpole's shadow, in conjunction with a simple trigonometric function and given the angle of the sun's rays in relation to the surface which the flagpole is planted in, no one would say that the length of the flagpole's shadow explains the height of the flagpole, although this derivation satisfies Hempel's criteria for a good (causal) explanation (see Hempel, 1965).

Bromberger's flagpole example, however, is not something which Hempel was unaware of. In Hempel (1965), he discussed a similar example concerning pendulums. The pendulum example is the following: we usually agree that the length of a pendulum l, in conjunction with the formula $t = 2\pi\sqrt{(l/g)}$, explains the period of the pendulum t, but disagree that the length of the pendulum is explained by its period in conjunction with the same formula, although both derivations are equally valid. For Hempel, the intuition which stays behind our inclination to assign different values to these two explanations is prompted by the belief that we can directly manipulate l, but do not do that with t. But this intuition is misleading, Hempel says, as in fact we cannot manipulate l independently of t—we always manipulate them together. Thus, according to him, "the common-sense conception of explanation appears to provide no clear grounds on which to decide whether a given argument that deductively subsumes an occurrence under laws is to qualify as an explanation" (Hempel, 1965, p. 353).

The prevailing opinion (see e.g. Psillos, 2007) is that Hempel's verdict on our explanatory intuition is not convincing. But to say merely that Hempel's argument is not convincing is not convincing either. One should explain why in such cases it is more safe to rely on what our intuition is telling us about the goodness of a given explanation than on what the DN model tells us about its goodness. The inferential content of the tentative explanations allows us to account for the explanatory asymmetry which the DN model is not sensitive to. Both the explanation of the length of the flagpole's shadow by the height of the flagpole and the explanation of the height of the flagpole by the length of its shadow allow for similar counterfactual inferences: if we manipulate the flagpole's height, and keep the angle of the light rays in relation to the surface which the flagpole is planted in constant, this will result in a change of the length of the flagpole's shadow. Similarly, if we change the length of the flagpole's shadow, that will lead to a change of the flagpole's height. The latter inference, however, is impractical insofar as we cannot manipulate the shadow directly. The only way to change it (while keeping the same angle of the sun's rays to the surface) is by varying the height of the pole. That is why we accept as explanatory the inference of the length of the flagpole's shadow from the flagpole's height, but the reverse inference is viewed by most of us as non-explanatory. The conclusion which is to be made from the proposed analysis is that the existence of extra-inferences allowed by a given explanation is not a sufficient condition for claiming this explanation a good one. It is to be added that at least some of the allowed new inferences should be useful.

Several objections could be raised against the proposal to evaluate the explanations' inferential content in order to assess their goodness. The first is connected to the question "What are the merits of a proposal whose realization finally leads to conclusions which our intuition has

[4] See e.g. (Fetzer, 2014), (Strevens, 2008), (Psillos, 2007). However, the text which is usually referred to (Bromberger, 1966) contains a different example (about the Empire State Building) having the same meaning (see Bromberger, 1966, pp. 92-93). As Wesley Salmon noticed (see Salmon, 1989, p.189), Bromberger probably never published the flagpole version of his counterexample, but it is this version that became the most popular.

already suggested to us?" The reply to this question is that there are cases where our intuitions split, and in these cases the proposal to assess the explanations' inferential content can help to decide between one of the alternative interpretations (see section below on circular explanations).

Another expected objection builds on the question "Is it not enough to take a pragmatic stance toward explanation, i.e. to evaluate directly the usefulness of tentative explanations instead of undertaking an epistemic analysis of the explanations' inferential content, which—as has been shown—should be followed by an evaluation of the usefulness of the extra-inferences allowed by the assessed explanation?" This is an interesting and important question which deserves a separate discussion. At the moment, it suffices to say that the main virtue of the proposed epistemic analysis lies in the fact that it reveals that an explanation could not be good from a pragmatic point of view if it is not epistemically good, i.e. if it does not allow for any extra-inferences to be drawn from it and the available knowledge related to the *explanandum.* But if the pragmatic virtues of explanations are always grounded in certain epistemic virtues, the study of the latter becomes important even from a pragmatic point of view.

The conjunction problem

Why is the conjunction of two or more *explananda* not conceived as a proper *explanans* for any of these *explananda*, even though they are entailed by this conjunction? The fact that the relation of entailment between the *explanans* and the *explanandum* is not always a sign of a good explanation was recognized as a problem for the DN model of explanation by Hempel himself (Hempel & Oppenheim, 1948). This problem was further discussed by Michael Friedman (1974), who allegedly proposed a criterion for distinguishing explanatory from non-explanatory conjunctions of laws. However, soon after that Kitcher (1976) showed that Friedman's criterion rules out many genuine explanatory conjunctions along with the bad ones. A recently proposed solution to the conjunction problem (Bangu, 2016) seems to be open to counterexamples which are similar to those that were raised against Friedman's solution, and like Friedman's solution, it works only if the *explananda* are laws. At the same time, it can be demonstrated that the evaluation of the inferential content of conjunctive explanations suggests better and simpler hints to distinguish between good and bad explanatory conjunctions. In addition, the analysis of the inferential content is not restricted to particular kinds of *explananda*. Look at the following tentative conjunctive explanation (1):

(1) "The lid conducts electricity" because "The lid conducts electricity" & "The light in the room is bright."

Does (1) allows us to learn something new related to the fact "The lid conducts electricity"? The answer is "No, it does not." Compare this "explanation" with the following one:

(2) "The lid conducts electricity" because "The lid is made of metal" & "All metals conduct electricity."

What is the difference between (1) and (2)? There seems to be no formal difference: in both cases the sentence "The lid conducts electricity" follows deductively from the conjunction presented after "because." Why then do we see the conjunction in (1) as non-explanatory and the conjunction in (2) as explanatory? Let us compare the inferential contents of (1) and (2). As we stressed above, (1) does not allow us to add anything to the fact "The lid conducts electricity," which is why it does not enhance our understanding of this fact. At the same time, (2) allows us to rule out some alternative explanations: knowing that the lid conducts electricity because it is made of metal, we are entitled to exclude the hypothesis that the lid conducts electricity because it is wet.

Again, one may try to disparage the analysis of the inferential content by saying that we do not need such an analysis because the same conclusion that (1) is non-explanatory could be drawn from the statement that (1) is circular (insofar as the *explanans* contains the *explanan-*

dum). However, as we shall see in the next section, not all explanations in which the *explanans* and the *exlanandum* overlap are (or should be) ruled out as bad explanations.

How to tell the difference between good and bad explanations when they appear to be circular

It has been well documented in psychology that children, non-expert adults, and scientists all face problems with the recognition of circular explanations (Rips, 2002; Baum, Danovitch & Keil, 2008; Hahn, 2011). The problems reported have multiple manifestations varying from a lack of sensitivity to circularity (in children and laypersons) to overstatements about certain circular explanations (in scientists). So far, these problems have received little attention in philosophy, although they have been broadly discussed among psychologists (see e.g. Hahn, 2011). As we shall see, there are good reasons to argue that both the insensitivity to circularity and the overstatements about circular explanations stem from the same facts that (a) most people, including most philosophers of science, assess explanations on the basis of their structure (the relation between the *explanans* and the *explanandum*); and (b) the structure of explanations does not provide hints to the difference between good and bad circular explanations. As we shall see, the value of seemingly circular explanations is better evaluated by their inferential content: good explanations, even if they are circular according to certain criteria, allow for new non-trivial inferences to be drawn, while bad circular explanations have poor inferential content.

Consider the following examples:

(3) "Opium induces sleep because of its dormative power."[5]

(4) "Hugo reacted violently because of his aggressive personality."

[5] This "explanation" was made famous by Molière ("*The Imaginary Invalid*," 1673).

For some people, including some psychologists (Bandura, 1999; Boag, 2011), there is no significant difference between (3) and (4); both are circular explanations. Many personality psychologists, however (McCrae & Costa, 1995), would disagree. They would argue (and in fact have argued) that (4) and similar personality trait explanations are not circular because they provide "surplus meaning," i.e. they "allow inferences which go beyond the observed data" (McCrae & Costa, 1995, p. 244), and it is this "surplus meaning" that makes them good explanations.

Indeed, on the one hand, explanations (3) and (4) have the same structure: in both the definition of the concept which is central for the *explanans* (the definition of "dormative power" in (3) and the definition of "aggressive personality" in (4)) involves the *explanandum.* "Dormative power" is defined in terms of inducing sleep, and the definition of "aggressive personality" includes references to various manifestations of violent behavior. Because of the difficulty in distinguishing the difference between (3) and (4) by looking only on their structure, most people either take both explanations to be equally acceptable (children, laypersons) or view them as equally bad (critics of the personality trait theory in psychology).

However, if one examines the inferential content of (3) and (4), the difference between them becomes much more salient. What can one infer from (3)? The only option is to draw the generalization "substances which induce sleep have dormative power" or any of its logical equivalents and implications. However, the knowledge that Hugo has an aggressive personality allows for a lot of non-trivial predictions. Knowing about Hugo's aggressive personality, we can expect that he would readily attribute hostile intentions to those who would eventually provoke him, and that he will not hesitate to cause harm if he sees it as a way to achieve his goals. These and other inferences which our knowledge about aggression as a personality trait allows us to draw are taken to constitute what McCrae and Costa called "surplus meaning" of personality trait explanations.

4. The proposal to assess the explanations' inferential content reveals a new perspective on the unification of the existing approaches to explanation

The existing plurality of approaches to explanation in contemporary philosophy of science is assessed differently by different philosophers of science.

The majority of the philosophers of explanation today have assumed straightforwardly that the different models of explanation which are discussed in the field represent different types of explanations used in real scientific practice (see Woodward, 2014). Some recent findings in experimental philosophy and cognitive science have backed up this stance by showing that there is a correspondence between the philosophical explications of explanation proposed in philosophy of science and the explanatory structures used by children and non-expert adults (Lombrozo, 2013; Colombo, 2016). Those who take seriously the existence of different types of explanation usually argue that it is both unnecessary and impossible to propose a general account which works equally well for all types of explanation. This group of philosophers share the view that it is better to focus on studying the specific features and the scope of application of the different types of explanation instead of looking for a general account which dismisses important characteristics of the subspecies of explanation.

Another group of philosophers, however, insist that the refusal to look for a general account of explanation leaves in the dark the question of why different types of explanation are explanatory. Some of these philosophers have launched different projects dedicated to the search for a unifying account of all (or at least some) types of explanation. A recent proposal, for example, suggests a counterfactual account that subsumes both causal and non-causal explanations (Reutlinger, 2016). This project seems quite successful in solving the unification task, but it faces serious problems in distinguishing between good and bad explanations. For example, the counterfactual account cannot tell us what is wrong with the big bang explanation of cultural differences in color preferences (discussed in section 2 of this paper). Insofar as a general theory of explanation would be essentially incomplete if it fails to tell us what makes a good explanation, the counterfactual account cannot be viewed as a successful general theory of explanation.

The proposal to assess the inferential content of explanations reveals a different perspective on the unification of the diverse approaches of explanation. Being compatible with all formal or semi-formal construals of explanation, including the counterfactual construal of causal and non-causal explanations, the assessment of the explanations' inferential content could be viewed as a universal tool for distinguishing good from bad explanations without considering the *explanans-explanandum* structure.

5. Conclusions

The main idea which has been introduced and defended in this paper builds on two widespread views: that (a) good explanations enhance our understanding of explained phenomena, and (b) we assess the understanding of a particular phenomenon on the basis of the inferences one is able to draw from his/her knowledge about this phenomenon. It has been shown that taken together, these two widespread views suggest evaluating the value of explanations by analyzing their inferential content. The inferential content of a given explanation is identified with the extra-inferences related to the *explanandum* which this explanation allows for. It has been shown how the evaluation of the explanations' inferential content helps to distinguish between good and bad instantiations of existing explicatory models of explanation. The evaluation of the inferential content tells us, for example, the similarity of the big bang explanation of cultural differences of color preferences, the explanation of the flagpole's height by its shadow's length, the explanations which use as an *explanans* the conjunction of several *explananda*, and the explanation of opium's effect of inducing sleep through its "dormative power." Analysis of the inferential content of these explanations reveals that they all have poor inferential content. As this proposal does not compete with existing explicatory models of the structure of explanation, but rather provides an important supplement to them, the proposal to assess the goodness of explanations on the basis of their inferential content reveals a new perspective on the unification of existing approaches to explanation.

However, it should be stressed that many important questions related to the proposal to assess the inferential content of explanations have not yet been answered. One of these questions is how this proposal relates to pragmatic approaches to explanation. What does it add to them, and how do they benefit from it? Another big question is to what extent the evaluation of explanations on the basis of their inferential content is present in one way or another in actual scientific practice. It has been shown in section 3 that personality psychologists have arrived at similar ideas which they have used to defend personality trait explanations against the circularity claims. However, it is both interesting and important to find out whether similar ideas have been promoted in other sciences.

References

Bandura A (1999). Social cognitive theory of personality. In *The Coherence of Personality: Social-cognitive Bases of Consistency, Variability, and Organization,* eds Cervone, D., Shoda, Y., 185–241. New York: Guilford Press.

Bangu S (2016). Scientific explanation and understanding: unificationism reconsidered. *European Journal for Philosophy of Science,* doi: 10.1007/s13194-016-0148-y.

Baum L A, Danovitch J H, Keil F C (2008) Children’s sensitivity to circular explanations. *Journal of Experimental Child Psychology,* 100: 146–155.

Baumberger G, Beisbart C, Brun G (2017) What is understanding? An overview of recent debates in epistemology and philosophy of science. In *Explaining Understanding. New Perspectives from Epistemology and Philosophy of* Science, eds Grimm, S., Baumberger, G., Ammon, S. New York: Routledge.

Boag S (2011) Explanation in personality psychology: “Verbal magic” and the five-factor model. *Philosophical Psychology* 24: 223–243.

Bromberger S (1966) Why-questions. In *Mind and* Cosmos, ed. Colodny, R. Pittsburgh: University of Pittsburgh Press.

Bruner J (1957) Going beyond the information given. In *Contemporary Approaches to Cognition,* eds Bruner, J., Brunswik, E., Festinger, L., Heider, F., Muenzinger, K. F., Osgood, C. E., Rapaport, D., 69. Cambridge, MA: Harvard University Press.

Colombo M (2016) Experimental philosophy of explanation rising: The case for a plurality of concepts of explanation. *Cognitive Science* 40 (5).

de Regt H W (2009) The epistemic value of understanding. *Philosophy of Science,* 76: 585–597.

de Regt H W (2013) Understanding and explanation: living apart together? *Studies in History and Philosophy of Science Part A,* 44: 505–509.

de Regt H W, Leonelli S, Eigner K (eds) (2009) *Scientific Understanding: Philosophical Perspectives.* Pittsburgh: University of Pittsburgh Press.

Fetzer J (2014) Carl Hempel. In:.) *The Stanford Encyclopedia of Philosophy,* ed. Zalta, E. (Winter 2016 Edition) https://plato.stanford.edu/archices/win2016/entries/hempel/

Friedman M (1974) Explanation and scientific understanding. *The Journal of Philosophy* 71: 5–19.

Gervais R (2015). Mechanistic and non-mechanistic varieties of dynamical models in cognitive science: explanatory power, understanding, and the ‘mere description’ worry. *Synthese* 192 (1): 43–66.

Hahn U (2011) The problem of circularity in evidence, argument and explanation. *Perspectives on Psychological Science* 6: 172–182.

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

Hempel C (1965) *Aspects of Scientific Explanation.* New York: The Free Press.

Khalifa K (2012). Inaugurating understanding or repackaging explanation? *Philosophy of Science* 79(1): 15–37.

Khalifa K (2013). The role of explanation in understanding. *The British Journal for the Philosophy of Science* 64(1): 161–187.

Khalifa K (2015) EMU defended: reply to Newman (2014). *European Journal for Philosophy of Science* 5(3): 377–385.

Kitcher P (1976) Explanation, conjunction and unification. *Journal of Philosophy* 73: 207–212.

Kvanvig J (2009) The value of understanding. In *Epistemic Value,* eds Haddock, A., Millar, A., Pritchard, D. Oxford: Oxford University Press.

Lipton P (2004) *Inference to the Best Explanation*, Second Edition, London: Routledge.

Lombrozo T (2013) Explanation and abductive inference. In *The Oxford Handbook of Thinking and Reasoning,* eds Holyoak, K., Morrison, R., 260–276. Oxford: Oxford University Press.

McCrae R R, Costa P T (1995). Trait explanations in personality psychology. *European Journal of Personality* 9: 231–252.

Newman M (2014) EMU and inference: what the explanatory model of scientific understanding ignores. *European Journal for Philosophy of Science* 4(1): 55–74.

Petkov S (2015) Explanatory unificationism and conceptualization. *Synthese* 192(11), 3695–3717.

Psillos S (2007) Past and contemporary perspectives on explanation. In *Philosophy of Science: Focal Issues*, ed. Kuipers, T. Amsterdam: Elsevier.

Reutlinger A (2016) Is there a monist theory of causal and non-causal explanations? The counterfactual theory of scientific explanation. *Philosophy of science* 83(5): 733–745.

Rips I J (2002) Circular reasoning. *Cognitive Science* 26: 767–795.

Salmon W (1989) *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.

Strevens M (2008) *Depth.* Cambridge, MA: Harvard University Press.

Toulmin S (1972). *Human Understanding*. Oxford: Clarendon Press.

Von Wright G H (1971) *Explanation and Understanding*. London: Routledge & Kegan Paul.

Woodward J (2014) Scientific explanation. In *The Stanford encyclopedia of philosophy*, ed. E. N. Zalta (Winter 2014 Edition) URL = <https://plato.stanford.edu/archives/win2014/entries/scientific-explanation/>.

ON UNDERSTANDING THROUGH AGENT-BASED MODELS

Richard David-Rus
Institute of Anthropology Francisc I. Rainer, Romanian Academy
Bucuresti, Romania
e-mail: rusdavid@gmail.com

Abstract:
The aim of this paper is to argue that it is more plausible to approach understanding from a special type of model—the ABS/IBMs models—as a non-explanatory form following some suggestions advanced by Lipton. I will first look to the type of explanation that some authors claimed is disclosed by these models: Weisberg's analysis of IBMs in ecology and Grüne-Yanoff's analysis of the Anasazi model. I argue that their analyses fail to show that these models qualify as explanatory understandings. This brings us to Strevens' "simple view," which claims the existence of a correct explanation behind any understanding, and his strategy of dismissing the challenges posed by non-explanatory forms. I argue that this strategy incurs damaging costs on his view. In the last part we turn to Khalifa's critique on Lipton's proposals and argue that it is based on an unjustified construal of Lipton's framework. I show how Khalifa's "argumentative strategy" fails to establish the superiority of actual understanding over possible explanation.

Key words: understanding, explanation, agent-based models.

The aim of this paper is to examine a type of understanding offered by a special kind of model. These models were recently invoked in the debate on scientific understanding to support a particular position. I will show that it is rather more plausible to view it as a non-explanatory form of understanding than to try to reduce it to an explanatory form.

In pursuing this aim, we will start with a brief overview of the debate on understanding and how models are involved. In the next step I will introduce the type of models I am referring to and evaluate the chances of having an explanatory type of understanding. By advancing this proposal, I am recalling some of Lipton's proposals (Lipton 2009), and I will suggest some amendments that make them more suitable for our goal. In order to make this point, I will address the critiques advanced by Strevens (Strevens 2014) and Khalifa (Khalifa 2013) against the possibility of any sort of non-explanatory understanding. I will end the paper by sketching some possible future directions for investigation.

1. Scientific understanding

The topic of scientific understanding drew the attention of researchers in recent years, and the debate can be seen as polarized between two main views. The first is a reductive position that seeks to reduce understanding to the epistemology of explanation and deny the subject of understanding any interesting contribution to the philosophical agenda, as explicitly stated by Khalifa: "any philosophically relevant ideas about scientific understanding can be captured by philosophical ideas about the epistemology of scientific explanation without loss" (Khalifa 2012, p. 17). The second view supports the idea of novelty in approaching the subject of scientific understanding, and therefore its irreducibility to mere explanatory knowledge. In order to justify this position, advocates point to aspects that cannot be captured by the competing view, arguing for the need to take into account means that go beyond propositional knowledge. These include skills and abilities such as those involved in making use of different theories in explaining phenomena (de Regt 2009), or comprehension abilities that go beyond the simple semantic one (Newman 2014), or the sort of abilities involved in manipulating models, as for example the models that simulate emergent properties of complex systems (Wilkenfeld and Hellmann 2014).

Reference to model-based understanding was considered by both camps involved in the debate, though there is a difference in the role and the weight that models had in supporting each position. In the reductivist camp, Strevens takes much care in addressing an important feature of models: their inherent idealization. He focuses on the way idealizations provide and enhance scientific understanding, and he provides a defense of the reductivist position as he elaborates on this issue (Strevens 2016). He is responding to the non-reductivist critique which claims that models reveal the limitation of reductivism and the necessity of including other ingredients. From a non-reductivist point of view, some models provide clear examples of how understanding is generated by necessarily including skills and abilities. The recent critique advanced by Wilkenfeld and Hellmann (Wilkenfeld and Hellmann 2014) claims that it is not possible to gain understanding through some special types of models without engaging in direct manipulation of those models. We will now examine these special types of models, namely the computational models known in scientific literature as agent-based simulation (ABS) or individual-based models (IBM).

2. The ABS/IBM models

ABS/IBMs constitute a class of computational models for simulating the actions and interactions of autonomous agents or individuals whose behavior follows a set of simple rules and for assessing their effect on the system as a whole, usually by generating an emergent macro-pattern. We will use examples of individual-based models (IBM) taken from biology and ecology which simulate swarming behavior in termite colonies and bird-flocking.

In the first case, the simulation done by Roach and collaborators (Roach et al 2013) contains only four simple rules for the behavior of a termite: 1) Walk randomly until you encounter a grain; 2) If you have a grain, drop it. 3) If you do not have a grain, pick one up; 4) Repeat. The images below show time points during the simulation that illustrate how the patterns of debris emerge:

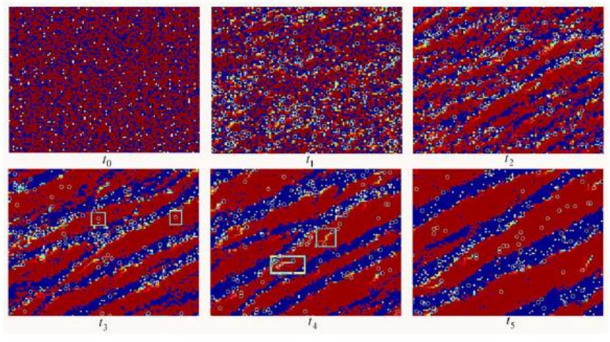

Figure 1 Six snapshots of termite swarm (Roach et al 2013)

In the case of the bird-flocking simulation (Boids model), there are just simple three rules: 1) Separate: steer away from your nearest neighbors to keep from getting too close; 2) Align: steer toward the average heading of your neighbors; 3) Cohere: move toward your neighbors. The emergent flocks appear as shown below using Wilensky implementation (Wilensky 1998):

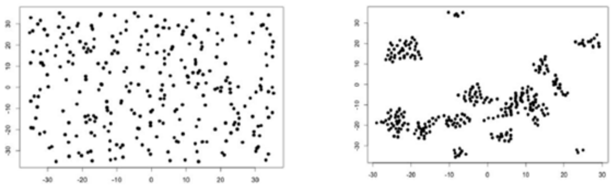

Figure 2 Initial random distribution (left) and formation of flocks (right) (Wilensky 1999).

The Anasazi simulation (Axtell et al 2002) is a more complex example taken from the social sciences, and involves simulating the population dynamics of the Pueblo Indians over a timespan of 500 years (800–1300AC), a group whose settlements were limited to the Long House Valley region of Arizona. The agents in the model are individual households, and their attributes are lifespan, movement capacities, nutrition requirements, and storage capacities. The model takes as input paleo-environmental data: meteorological, groundwater, and sediment deposition. These data determine the production of maize per hectare for different categories of farming land. Secondly, household movements (resettlement, land cultivation, procreation, annual deaths) are computed based on the previous results and the behavior rules. The behavioral rules of the agents are provided by local ethnographical and anthropological studies of historic Pueblo people and other subsistence

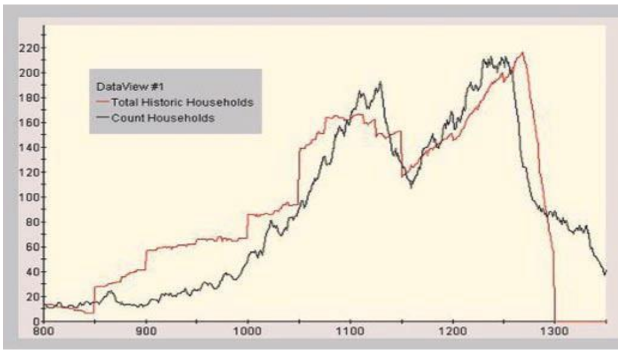

Figure 3 Households dynamics—historic and computed (Axtell et al 2002).

agriculturalists. The output gives us the features of the history of the settlement, including population ebb and flow, changing spatial settlement patterns, and eventual decline. Simulated population levels closely follow the historical trajectory and also reproduce the dynamics of the settlements clustering, as show in the figure 3

3. What kind of an explanation?

The first issue we will address is the possibility of identifying a certain type of explanation through such models. In Wilkenfeld and Hellmann's critical paper (Wilkenfeld and Hellmann 2014) of Strevens' account, both sides seem to take for granted the existence of an explanation behind the simulation. The authors' main criticism of Strevens' account—which restricts understanding to grasping explanations rendered through propositional knowledge—is that merely grasping the simple set of rules that describes the simulation is insufficient. But the critique does not identify any specific explanatory format that these rules should articulate. It is instead suggested that the format underlying Strevens' account is a causal explanation rendered through the deductive-nomological schema, with rules playing the role of lawlike claims and the boundary conditions taking the values of the parameters set in the simulation.

Such a view is not uncommon, and is more or less explicitly supported by other authors, especially scientists who assume that ABMs/IBMs usually track causal explanations. From the philosophers' perspective, the most commonly circulated construal is to render this causal sort of explanation in terms of Woodward's counterfactual manipulation account (Yllikoski 2014). We will discuss this approach as applied to our case in the second part of the paper, while addressing Khalifa's critique. For the moment, notice that despite the fact that this sort of approach captures important features of ABMs, such as the manipulation aspect, it leaves us with a major problem: the lack of any clear explanatory connection between the micro-level and the generated macro-pattern. We may therefore be justified in searching for an explanatory structure to illuminate this connection. In the following section we will discuss some proposals advanced by two different authors that identify such a structure behind the simulation. I will argue that that the success of both attempts is questionable.

4. Weisberg's analysis

The explanatory virtues of such computational models were discussed by Weisberg in a special paper on IBMs as applied to ecology (Weisberg 2014). He distinguishes between two sorts of explanation: one explaining the emergence of the phenomenon (which he calls *emergence explanation*) and the second explaining the emergent phenomenon, which he saw as a type of *mechanistic explanation*. The first is viewed as a kind of reductive explanation that shows how emergent phenomena arise from lower-level interactions, and cites the paradigmatic case of the explanation of phenomena described by thermodynamics through statistical physics. Weisberg dismisses the plausibility of this sort of explanation in the case of IBMs in ecology, since we are far from claiming even the slightest possibility of a theory as advanced as statistical mechanics "that would connect lower-level biotic and abiotic properties to higher-level ecological properties."

In case of the mechanistic explanation, Weisberg believes the situation is quite different. He claims that IBMs allow us to "develop a generalized mechanistic understanding of the dependence of higher-level properties and patterns on lower-level mechanistic factors." In making his point, Weisberg goes beyond the simple counterfactual manipulation involving just changing the simple initial rules and points to the possibility of extracting some major factors that are related though a causal graph. He employs a simulation from ecology: the simulation of the spatial pattern of pre-populated Europe's beech forest as developed by Rademacher and collaborators (Rademacher et al 2004). He argues that despite the variety and multiplicity of the variables and interactions involved in the model, most of the forest structure might be explained by only two main factors: vertical light competition and neighborhood interactions.

Following Weisberg's suggestion, we might see a way to identify the explanatory structure generating the understanding: the causal graph revealing the mechanistic explanation behind it. Nevertheless, as one might notice, Weisberg is switching to a richer and "more realistic" model (described by no less than sixteen parameters) from the initially simpler Boids model. In such cases as this or the simulation of swarming behavior, there are only a few rules in the model, a situation that makes the exploration of the parameter space and the extraction of an eventual causal graph unlikely or less informative. There is of course the option to claim the eventual enrichment of the model with more variables until we could get at to stage where we might extract such a graph. But this eludes the point that these few rules are sufficient to generate the macro-pattern, and that we are gaining a sort of understanding even in the cases of such simple models.

5. The Anasazi model

A different suggestion pointing to a non-causal sort of explanation was advanced by Grüne-Yanoff (Grüne-Yanoff 2009) in a detailed discussion of the explanatory virtues of the Anasazi model. After dismissing the different classical types of explanations (causal, possible causal histories), Grüne-Yanoff's proposal is to see the simulation as providing a potential functional explanation (best rendered though Cummins' variant of functional explanation). Such explanations are meant to show how lower-level capacities contribute to the system capacity, i.e. the capacity of a system being explained in terms of the capacities of the components it contains, and how they are organized. In the Anasazi case, the model "organizes these capacities in a specific 'program' (the behavioral rules of the households, the yield functions of the farming plots) so that their combined operation, when fed with meteorological and soil data, produce the population data" (Grüne-Yanoff 2009).

David-Rus (2012) critically analyzed Grüne-Yanoff's approach (David-Rus 2012) and argued that his solution remains unsatisfactory. For our current purpose, it suffices to say that such an explanation lacks validation and remains merely a potential explanation, rather than an actual one. It instead satisfies what Strevens calls the internal criteria for correctness, i.e. the narrow notion of understanding, which is distinct from the broader notion of satisfying the external criteria of correctness, i.e. the proper one providing genuine understanding-why of a phenomenon.

The above solution is also along the lines suggested by Grüne-Yanoff and Weirich in a synthetic paper reviewing the explanatory virtues of simulations (Grüne-Yanoff and Weirich 2010). They see simulations as providing us with *partial potential* explanations. The explanations are only partial since they can identify and isolate only some factors at work in the inquired phenomenon. They fall short in fulfilling the truth requirement by identifying and exposing *possible* but not actual mechanisms, and are therefore better characterized as potential explanations.

In the next section, we will appeal to the idea of non-explanatory understanding as a more adequate way to capture this sort of understanding. Potential explanation could be seen as providing such an understanding, and we will recall Lipton's view in order to introduce it. Taking this view as a starting point, I will suggest some additional amendments and necessary extensions in order to make it more plausible for our goals. But before getting to these suggestions, we will address the critiques made by Strevens and Khalifa against any non-explanatory form of understanding.

6. Non-explanatory understanding

In this section we will briefly recall Lipton's main suggestions on non-explanatory understanding, which we regard as adequate for application to our case. Peter Lipton (Lipton 2009) argues for the possibility of non-explanatory understanding by identifying the cognitive benefits of an explanation with understanding. He further argues that such benefits might also be gained through other means than explanation. The benefits are identified as knowledge of causality, necessity, possibility, and unity, through such diverse non-explanatory means as models, visualizations, manipulations, thought experiments, or possible explanations. In Lipton's view, visual models and manipulations of physical systems can provide tacit knowledge of causes and therefore causal information without explanation—e.g. by using a planetarium for retrograde motion. Optimization and symmetry arguments, as well as thought experiments, constitute examples of non-explanatory deductive inferences that usually provide knowledge of necessity—e.g., Galileo's argument for the independence of acceleration and mass. Tacit knowledge of unification can be provided through non-explanatory analogies, as illustrated by the exemplars mechanisms that Kuhn emphasizes, through the scientists' abilities to select problems and generate solutions. Finally, incorrect, false, or "merely potential" explanations can provide knowledge of possibilities, a situation that cab be seen in our case.

7. The correct explanation behind—Strevens' move

Strevens does not articulate a separate critique on non-explanatory understanding, but refers to it in the context of articulating his approach on scientific understanding and meeting the potential challenges posed by such forms. Strevens' approach—called the "simple view"—states that: "An individual has scientific understanding of a phenomenon just in case they grasp a correct scientific explanation of that phenomenon" (Strevens 2013). Grasping correct explanations involves grasping two things: that the state of affairs is represented by the propositions obtained (understanding-that) and that they instantiate the explanatory structure. The second part explicitly assumes the existence of an explanatory structure, and Strevens is drawing mainly on this point when trying to meet the challenges from non-explanatory forms. He is not offering a well-articulated response to these challenges, but is rather providing us with some suggestions. Behind them one might see the overall strategy that he is adopting and that we will try to address.

In his paper on scientific understanding, Strevens refers only to some of the particular forms of non-explanatory understanding suggested by Lipton. In case of understanding provided through implicit causal knowledge gained by using visual models or systems manipulation, his move is to claim that grasping might also apply to an implicit explanation. This means that it does not necessary imply that the person who grasps an explanation can articulate it in an explicit way. We see here his main strategy of stating that understanding might be gained through such non-explanatory methods only insofar as they can be seen as "reflections" of the grasping of a correct explanation, which is therefore assumed to exist in the background. Strevens also makes use of this strategy in examining other non-explanatory forms. He claims that it might be sufficient in some situations to grasp only parts of the correct explanation, or even just some "psychologically operative pre-theoretical principles" that are among the constitutive parts of a correct explanation. This is intended to meet the challenge from the case of Galileo's argument in Lipton's example, where the "illumination" comes from some sort of pre-theoretical grasp of physics.

The problem with this strategy is that it does not come without costs. These moves weaken his account,[1] threatening both components of grasping: the representational part and the instantiation of the explanatory structure. His suggestions require further substantiation that he is not offering in order to validate the approach. In the absence of a sufficiently detailed articulation, it is hard to precisely criticize his position, but we might still see the potential dangers that threaten his project. His move of allowing implicitness for explanations impacts on both parts of grasping, raising especially concerns related to the objective evaluation of explanations. The claim about the possibility to grasp only fragments of the correct explanation could bring about a sort of underdetermination argument invoking the case of same parts of an explanation being shared by many correct explanations; in this case one cannot determine which correct explanation is grasped. Other factors might be determinant in such cases, but this shows the need for more substantiation. As with implicitness, by allowing pre-theoretical intuitions to play a validating role, one runs into the issue of distinguishing between genuine and fake forms of understanding.

Now, even if his general strategy of dealing with non-explanatory forms of understanding is not satisfactory, one might invoke some other suggestions that Strevens is offering. Such are the ones concerning potential explanations and ones that refer to the distinction between understanding-why and understanding-with. In the case of potential explanation, Strevens acknowledges a different sort of understanding-why than in the case of an actual explanation. He draws a distinction between external and internal conditions that an explanation might fulfill, analogous to the logical and empirical requirements in Hempel's deductive-nomological account. Fulfilling just the internal conditions corresponds to having a potential explanation replacing the correct one in the simple view, providing a narrow notion of understanding. Strevens even defends this narrow notion as being not much closer to the purely phenomenal sense of understanding than the broader one. Though in contrast to the other ways of non-explanatory understanding from Lipton's set, for which Strevens insists on the existence of a correct explanation, he admits a different type of understanding in the case of potential explanation. Strevens explicitly articulates the above distinction in discussing understanding-with versus understanding-why, the first of these being suited to the case of models.

Understanding-with is a different species of understanding than understanding-why (as is understanding-that), and is not the main object of his analysis. He is therefore not offering a detailed account of understanding-with, nor its relation to the other two types of understanding; rather this species is introduced in order to clarify and answer some problematic issues raised by his account. From its sketchy characterization we are informed that the object of understanding-with is a theory (or a theoretical construct, i.e. a model or representation) rather than a phenomenon, and that it is primarily gained through grasping internally correct explanations. As Strevens specifies: "to have 'understanding with' of Newtonian physics is to be able to construct or grasp an array of Newtonian explanations that are good in the sense that they are internally correct" (Strevens 2013).

One might try to take this proposal to account for the sort of explanations rendered through a com-

[1] As Wilkenfeld and Hellmann also notice: "if, on the other hand, one broadens his notion of grasping to include more active abilities and weakens the requirement that explanations be construed as propositional, then nothing substantive remains of Strevens's thesis." (Wilkenfeld&Hellmann 2014, p. 46).

putational model providing an understanding-with model. Nevertheless, applying it to our examples is not unproblematic. In the IBM or ABS models, there is no internally correct array of explanations that can be grasped, as in case of Newtonian physics. Though one might claim that internal correctness is delivered through the iterated application of constitutive rules, there is no possible grasping of the multiple successive derivations, nor of the link between the micro-specifications and the emerging macro-pattern. The derivation of the macro-pattern from the interaction rules and boundary conditions is epistemically opaque. So it appears that we might not have in this case a sort of understanding-why as Strevens describes.

8. Non-explanatory understanding through potential/possible explanation

We will focus now on the very specific sort of non-explanatory understanding that might be claimed in IBM/ABS models: the potential/possible explanations. Lipton's discussion around understanding from possible explanations focuses especially on the way actual understanding is gained through such explanations, even false ones, and even in a radical way that involves more than merely sharing some parts of an actual explanation or approximating it. According to Lipton's view, potential explanations provide us understanding not only of a possible world, but of the actual one through the modal information they deliver. Such information might show how the actual mechanisms belong to a class of mechanisms, or they might reveal some contingency of the mechanism or some necessity. They can improve our understanding on "actual capacities, powers and causal relations" (Lipton 2009, p. 51). Even when they fail to describe the actual mechanisms, we might gain "oblique information" on the phenomenon's occurrence, "learning that a particular way it could have occurred was not the way it did occur" (Lipton 2009, p. 51).

Before turning to the critique that Khalifa articulates against Lipton's proposal, I would like to add some amendments on Lipton's approach. One concerns the fact that he relies on the distinction between actual and potential characterization of explanation, applying it to understanding as well. This qualification appears to me to be less justified in the case of understanding, in light of a more general investigation of its non-explanatory forms. The distinction retains the unnecessary reference to explanation, failing to clearly decouple explanation from understanding—from the actual explanation in this case—and makes the approach vulnerable to reductive objections such as Khalifa's critique.

The need for decoupling becomes clearer when it is seen that the contribution of a possible explanation to the understanding of an inquired phenomenon does not necessary involve reference to an actual explanation. Nor is the existence of an actual explanation guaranteed in such a case. Such an understanding might also bring information unrelated to the actual mechanism or to its qualification. This sort of situation is not addressed in Lipton's argumentation. He is rather focusing on the particular issue of how possible explanations offer actual understanding by providing modal information on the actual explanatory mechanism. Modal information on the actual mechanism is one benefit, but not the single one delivered in this case. Using Lipton's way of specifying the benefits, we could also extract knowledge in terms of causality, i.e. identifying causal factors at work in this situation, or on their necessity, since some of them are indispensable in generating the macro-pattern. In the last section, I will suggest other ways of extending the benefits beyond Lipton's way of approaching them.

9. Khalifa's critique on Lipton's proposal

In his paper "The role of explanation in understanding" (Khalifa 2013), Khalifa articulates the most explicit critique of Lipton's proposal. His paper aims to establish and defend what he calls "explanatory idealism," claiming that explanation plays a special role as an ideal of understanding, and therefore any other modes of understanding (non-explanatory included) "ought to be assessed by how well they replicate the understanding provided by good and correct explanation" (Khalifa 2013). Khalifa proceeds in the first section to present Lipton's framework and to articulate a clarification for it. In the third part he presents explanatory idealism, and offers us also a "general argumentative strategy," as he calls it, for how to interpret Lipton's examples. In the rest of the paper he applies this strategy to the particular cases that Lipton discusses; we will be interested in the one related to possible explanation and modal information .

In his presentation of Lipton's framework and its subsequent clarification, Khalifa forces a reading that imposes, in my view, an unfair construal. This move is presented at the beginning in what he calls Lipton Assumption (LA); the rest of his argumentation could be seen as an unfolding of this initial reading. LA is a restatement of Lipton's claim that "it is more natural to identify understanding with the cognitive benefits that an explanation provides rather than with the explanation itself" (Lipton 2009, p. 43), and in a first reformulation appears as "If knowing that b constitutes understanding of p, then a correct explanation e of p provides knowledge that b." But Khalifa offers us a further "friendly articulation" of LA that constitutes the basis of his argumentation: "If knowing that b constitutes understanding of p, then *there exists*[2] a correct explanation e of p such that knowing that e explains p entails knowing that b." In this way he extracts a strong existential claim about explanation. This move appears to be problematic, since such a strong reading can hardly be justified. The strong reading actually contradicts Lipton's further move and the main intention of his argument, which implies the possibility of obtaining these benefits without an explanation, and in this way decoupling explanation from understanding. In Lipton's words, "by distinguishing explanations from the understanding they provide, we make room for the possibility that understanding may also arise in other ways" (Lipton 2009, p. 44). By denying this move and imposing the strong existential reading one might claim, as other author has,[3] that Khalifa is missing the point of Lipton's proposal.

In the next section Khalifa tries to justify "explanatory idealism" by claiming that it provides a "plausible rationale" for the privileged role explanation is given in LA. But this already assumes his strong reading. By further providing some examples of "misunderstandings without explanation," he intends to strengthen his position by pointing to the asymmetries between genuine non-explanatory forms that follow the ideal standard and misunderstandings, i.e. deviant non-explanatory forms. Though he acknowledges that "*some* kind of understanding may still be in play" (Khalifa 2013, p. 8), even in such cases he insists that the benefits provided through these means should not be identified with understanding-why.

In the following section, Khalfia proceeds to set the way Lipton's examples should be interpreted by stating the SET thesis (Superior Explanation Thesis): "For any non-explanatory way, w, to understand p, there exists a correct and reasonably good explanation, e, such that the understanding of p provided by w is *a proper subset*[4] of the understanding of p provided by knowing that e" (Khalifa 2013, p. 9). In order to apply this thesis and unpack the way in which one understanding is a subset of another, he offers us a "general argumentative strategy" to measure "degrees of understanding." It mainly states that a subject who gains understanding through a correct and "reasonably good" explanation has access to all the benefits of understanding, including the ones provided though a particular non-explanatory way w (so that another subject who gains understanding only through w has less understanding that the first subject). This strategy is further applied in each case of Lipton's examples. Our discussion will be limited to the case of possible explanations.

10. Understanding through potential/possible explanations

Khalifa proceeds to implement his strategy in this case by showing that the understanding from a correct explanation is "greater" than the understanding from a potential/possible explanation. He acknowledges that one might gain some sort of understanding through a possible explanation, but his argument aims to show that it is inferior to understanding from an actual explanation. In order to accomplish this, he appeals to a particular analysis of causal explanation proposed by Woodward (Woodward 2005). According to Woodward, an explanation should be able to answer w-questions, i.e. "what-if-things-had-been-different" questions. The quantitative measure is thus rendered in terms of the number of counterfactual inferences that constitute the answers to the w-questions. Khalifa proceeds to argue that the answers to w-questions from a possible explanation are a subset of the ones from a correct explanation.

In evaluating Khalifa's critique, we might first notice that the appeal to Woodward's approach assumes the restriction to causal explanations, as it was designed primarily to account for this type of explanation. The extension to other types of explanation might be questionable. Though in the discussed examples the causal pattern is a plausible one, the range of possible explanation also covers other types of explanations. In this sense his critique is a limited one, leaving out all types of possible explanations other than causal ones.

Considering now Khalifa's "argumentative strategy," the problem arises in his second step of stating that a subject S* who knows the actual explanation also knows "all of the answers to the w-questions that comprise S's understanding of p via a potential explanation" (Khalifa 2013, p. 11), and therefore S* understands more than S. It seems questionable to claim that there is always a strict inclusion relation between the counterfactual inferences of the two explanations. It might be possible to have

[2] Italics are my emphasis.
[3] "In my view, however, Khalifa misses the main point of Lipton's argument" (de Regt 2013, p. 507)
[4] Italics are my emphasis.

a non-empty intersection of the set of answers to w-questions, but a strict inclusion is an unjustified claim. The reason is based primarily on the fact that a potential explanation usually gives a different insight by simply bringing or making relevant different information than the one from the actual explanation. Answering w-questions involves varying initial conditions with reference to the given explanation, i.e. explanans, and see how they affect the occurrence of explanandum. As the explanatory information is different, there will be a different set of counterfactual inferences as answers to w-questions.

One might claim that Khalifa intends to answer some variant of these worries when he addresses the suggestion that it is not entirely justified to claim that explanatory knowledge entails knowledge of all sorts of specific possibilities. His reply states that "while it is certainly true that not all good and correct explanations answer every w-question, SET only requires one correct explanation to answer the same w-questions as a merely potential explanation" (Khalifa 2013, p. 12). This answer misses the point of the critique, and in my view fails to address it. Moreover, by claiming that one may pick one correct explanation from a set of "all good and correct explanations," a radical change is made in the initial setting of the problem that refers to one actual explanation and the plurality of possible explanations.

In addition, understanding from potential explanation simply implies other consequences in terms of configuring knowledge. If understanding is more than knowledge, as many authors suggest—implying a sense of how facts fit together (Kosso 2007), i.e. a global, "more holistic" modality of configuring the information—then potential explanation configures knowledge in a different way than an actual explanation, or contributes in other ways to the general sense of the overall coherence; this fact is reflected in the set of answers to w-questions, i.e. the counterfactual inferences that could be drawn in the case of each explanation.

A last point that should be highlighted is related to the insufficiency of having a quantitative criterion limited to the number of counterfactual inferences for measuring and comparing understanding, especially when non-explanatory forms are also involved. In order to advance the investigation, one also needs to consider the relevance of these inferences for expanding knowledge and their fruitfulness for opening new methods of inquiry. As de Regt & Gijsbers (de Regt & Gijsbers 2016) recently argued, understanding is also linked to scientific effectiveness and scientific success,[5] and "the tendency to produce useful scientific outcomes such as correct predictions, successful practical applications and fruitful ideas for further research" (de Regt & Gijsbers 2016). It is linked to possible routes to be followed in advancing our knowledge, fruitfulness in generating and exploring other scenarios and hypotheses, and the ability to articulate such scenarios and produce alternative explanations. Depending on the explanation considered, one could translate it in terms of the plausibility of the consequences drawn from the information at hand. Understanding from different explanations will enhance the plausibility of following some specific routes, will direct the inquirer to some specific tracks to follow, and raise specific new research questions. This sort of fruitfulness of an explanation is not accounted for in a limited quantitative criterion for defining understanding, as the one previously discussed is.[6]

11. Lipton's example of a rigged boxing match

We will further articulate our critique by using one of the examples that Lipton advanced and that Khalifa also discusses: a rigged boxing match. A boxing match between Able (A) and Baker (B), in which B is a far better boxer than A, is rigged so that B will take a dive in the 10th round. In Lipton's view, knowing this fact "helps us to understand why Able won, even if as a matter of fact Able floors Baker with a lucky uppercut in the fifth" (Lipton 2009, p. 51). Khalifa disagrees and restates his view on the superiority of understanding from actual explanation in this case: "a person who only knew that Baker could have lost the match because he would have taken a dive in the tenth, but does not know that Baker was actually knocked out fair and square in the fifth, doesn't understand why Baker lost the match as well as someone who also knows about the actual course of events" (Khalifa 2013, p. 12). Khalifa does not spell out in more detail his "argumentative strategy," so much is left implicit. This would have revealed the points where his strategy fails, and where some questionable moves become evident. Instead he is just restating his point regarding the

[5] Their intention is to replace the criterion of representational veridicality of understanding with an effectiveness condition.

[6] One might claim that this could be captured in the sort of quantitative measure mentioned, but it is less plausible that it could be reduced to such a local measure as the number of answers to w-questions from a specific explanation in a particular context of inquiry.

superiority of understanding from actual explanation.

There is indeed a difference in understanding in relying on the two different explanans, but the fact that one would be greater than the other is less clear. Recasting the comparison in terms of an inclusion relation between sets of counterfactual inferences is highly questionable. The understanding from the actual explanation does not contain the understanding from the possible explanation unless the person also knows about the rigged match, and so it includes extra knowledge which goes beyond the actual explanation. Knowing the actual explanation, i.e. the lucky uppercut, does not imply knowledge of the rigging, and so the counterfactual inferences generated around this fact are not subsumed under the set from the actual explanation. Khalifa unjustifiably assumes this extra knowledge, and it is not clear from his answer to the above-mentioned objection that there is another, correct explanation (from all the good and correct ones) that applies in this case. Besides, based only on the actual explanation, the answer to what would have happened if A did not get the lucky uppercut is that he probably would have lost. So it fails to get the counterfactual situation right.

We turn to the point that understanding should be also considered from the perspective of its fruitfulness. In our case, depending on the explanans we use in order to gain understanding (from either a potential or actual explanation), we open new perspectives and routes of advancing our inquiry and knowledge. The plausibility of some downstream inferences are different in the case of each explanans: understanding that A won because the match was rigged raises the plausibility of different scenarios and ways to expand knowledge from the case in which I see the winning result as the consequence of the lucky uppercut. In the first case, considerations that rise in plausibility might be linked to rigging practices in the boxing world, morality of the people involved, unfairness, etc., while in the second case, the qualification "by chance" appears less illuminating when considering opportunities for further knowledge-expansion.[7] Moreover, it might be the case, depending on the context, that by understanding that the match was rigged, one might gain greater understanding of the situation than just by pointing to bare "chance." Contrary to Khalifa's view, the understanding through the potential explanation might be more consistent in such a case, even if seen only in terms of fruitfulness and knowledge advancement.

12. Perspectives for a working agenda

Before concluding, I would like to point on a few of the more urgent aspects that should be the subjects of further inquiry as a consequence of this discussion. I will suggest a possible working agenda by sketching some directions for investigating non-explanatory ways of understanding. I will insist just on two main points.

The discussion here suggested that there are no good reasons to reject non-explanatory forms of understanding as characterizing the understanding provided by ABMs. As a consequence, instead of directing our efforts toward searching for an explanatory pattern behind understanding, or an ideal correct explanation that would validate this understanding, we would do better to concentrate on investigating the benefits provided by such models. The benefits that Lipton discusses constitute just a limited set, as other authors have noted (Khalifa among them), showing the need to further expand the range. For example, the analysis offered by Yllikoski (Ylikoski 2014) specifically addressed the benefits of using ABS in sociological research, though I do not agree with his working assumption referring to "explanatory understanding" and identifying understanding with the number of *what-if* inferences. Yet his characterization of the three ways in which ABS increases our understanding is inspiring: they force "theorists to be more explicit about their assumptions," make "inferences more reliable," and expand "the number of *what-if* inferences that one is able to make." (Ylikoski 2014, p. 329) I do not see any good reason not to consider these benefits as part of a non-explanatory form of understanding.

This brings us to the second point I would like to emphasize: the neglected issue of the relation between non-explanatory forms of understanding and explanatory ones. We should be examining the relation between the benefits provided through non-explanatory forms and the fully-articulated explanations, and their role in the processes of generating, applying, and modifying explanations. One way to pursue such an investigation could be to reframe the issue in a modeling context and make use of the recent general account on modeling practices provided by Weisberg (Weisberg 2013), which provides a general frame for inquiry. One might notice that in his approach, the reference to understanding takes precedence over references to explanation and explanatory episodes, making it a suitable frame for an investigation of this kind.

[7] The context of inquiry could be the determining factor in this case.

13. Conclusions

This paper aimed to argue that such models as IBM or ABS could be regarded as offering a sort of non-explanatory understanding, and in this way better advance our research into the contributions made by such models. I tried to show that the critiques expressed by some adherents of the reductive view of understanding are insufficient to dismiss a philosophical investigation into non-explanatory forms of understanding. I believe that there is a need to engage in further inquiry regarding the different cognitive benefits that models can offer, and into the relations between such non-explanatory forms of understanding and the explanatory forms.

References

Axtell R L, Epstein J M, Dean J S, et al. (2002) Population growth and collapse in a multiagent model of the Kayenta Anasazi in Long House Valley. *Proceedings of the National Academy of Sciences* 99: 7275–7279

David-Rus R (2012) Explaining by using artificial societies. *European Journal of Science and Theology* 8(3): 103–113

de Regt, H W, Leonelli, S, Eigner, K (2009) Focusing on scientific understanding. In *Scientific Understanding: Philosophical Perspectives*, eds de Regt H W, Leonelli S, Eigner K, 1–20. Pittsburgh: University of Pittsburgh Press.

de Regt H W (2013) Understanding and explanation: Living apart together?. *Studies in History and Philosophy of Science Part A* 44(3): 505–509

de Regt H W, Gijsbers V (2016) How false theories can yield genuine understanding. In: *Explaining understanding: New perspectives from epistemology and philosophy of science,* eds Stephen G, Baumberger C, Ammon S. (ed). London: Routledge.

Grüne-Yanoff T (2009) The explanatory potential of artificial societies. *Synthese* 169: 539–555

Grüne-Yanoff T, Weirich P (2010) The Philosophy and Epistemology of Simulation: A Review. *Simulation & Gaming* 41: 20 –50

Khalifa K (2012) Inaugurating Understanding or Repackaging Explanation? *Philosophy of Science* 79*:* 15–37

Khalifa K (2013) The Role of Explanation in Understanding. *The British Journal for the Philosophy of Science* 64: 161–187

Kosso P (2007) Scientific Understanding. *Foundations of Science* 12: 173–88

Lipton P (2009) Understanding without explanation. In *Scientific Understanding: Philosophical Perspectives*, eds de Regt H W, Leonelli S, Eigner K, 43–63. Pittsburgh: University of Pittsburgh Press.

Newman M. (2014. EMU and inference: what the explanatory model of scientific understanding ignores. *European Journal for Philosophy of Science* 4(1): 55–74.

Rademacher C, Neuert C, Grundmann V et al. (2004) Reconstructing Spatiotemporal Dynamics of Central European Natural Beech Forests: The Rule-Based Forest Model Before. *Forest Ecology and Management* 194(1): 349–68

Roach J, Ewert W, Marks R, Thompson B B (2013) Unexpected emergent behaviors from elementary swarms. In: System theory (SSST), 2013 45th Southeastern Symposium on, Baylor University, Waco, TX, USA, March 11, 2013, 41–50

Strevens M (2013) No understanding without explanation. *Studies in History and Philosophy of Science Part A* 44(3): 510–515

Strevens M (2016). How Idealizations Provide Understanding. Forthcoming in *Explaining understanding: New perspectives from epistemology and philosophy of science,* eds Stephen G, Baumberger C, Ammon S. London: Routledge.

Weisberg M (2012) *Simulation and similarity: Using models to understand the world*. New York: Oxford University Press.

Weisberg M (2014) Understanding the Emergence of Population Behavior in Individual-Based Models. *Philosophy of Science* 81(5): 785–797.

Wilkenfeld D A, Hellmann J K (2014) Understanding beyond grasping propositions: A discussion of chess and fish. *Studies in History and Philosophy of Science A* 48: 46–51

Wilensky U (1998) NetLogo Flocking model Available via Center for Connected Learning and Computer-Based Modeling Northwestern University Evanston IL http://ccl.northwestern.edu/netlogo/models/Flocking

Wilensky U (1999) NetLogo available via Center for Connected Learning and Computer-Based Modeling Northwestern University Evanston IL http://ccl.northwestern.edu/netlogo

Woodward J (2005) *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Ylikoski P (2014) Agent-Based Simulation and Sociological Understanding. *Perspectives on Science* 22: 318–335

CAUSATION AND EXPLANATION IN PHENOTYPE RESEARCH

Özlem Yılmaz
Sabancı University, Faculty of Engineering and Natural Sciences
Tuzla, İstanbul, Turkey
e-mail: ozlemyilmaz@sabanciuniv.edu

Abstract:

A phenome occurs through the many pathways of the complex net of interaction between the phenome and its environment; therefore researching and understanding how it arises requires investigation into many possible causes that are in constant interaction with each other. The most comprehensive investigations in biology are the ones in which many biologists from different sub-areas—evolutionary biology, developmental biology, molecular biology, physiology, genetics, epigenetics, ecology—have collaborated. Still, biologists do not always need to collaborate or look for the most comprehensive explanations. A more standard investigation in biology occurs within a single subarea, and uses well-defined experiments with very specific conditions. This paper is about causation and related explanation in plant phenome research and its relevance to Aristotle's Theory of Four Causes. I argue that there are causes which resemble Aristotle's formal, material, and efficient causes in phenotype explanation and occurrence; but causes which resemble Aristotle's final causes occur in phenotype explanation only, not in the occurrence.

Key words: Plant Phenome, Causation, Explanation, Complexity, Interaction, Four Causes.

1. Introduction

There are many factors in phenome occurrence, and these are related to each other in such a way that they are constantly affecting each other through a complex net of many processes. This interrelated and complex way of existing together causes the phenotype. Understanding the complex pathways of the interaction net between genotype, phenotype, and environment requires investigation into many possible causes, including ecological, physiological, evolutionary, developmental, molecular, epigenetic, and genetic factors. When scientists reach an explanation of a phenotypic trait, this explanation concerns one of these factors, so they are giving an explanation of one of the phenome's parts. If they want to give a comprehensive explanation of a phenotypic trait, then research into *all* of these factors is required, which likely necessitates collaboration with other scientists. The completeness of this explanation is of course limited by our current scientific knowledge, unless we are asserting new laws.

I think biology as a branch of Natural Sciences has its own specific features and it is different from physics or chemistry in many ways. This is not the subject of this paper, but I would like to mention that I believe these differences are related to the great complexity in phenome occurrence. Biologists usually do very well defined and strictly confined experiments after which they give explanations of some very specific parts of very specific biological phenomena. They do not often conclude generalizations as physicists or chemists usually do. Clearly this does not mean that biology is less powerful than them, it only means that it is a bit different from them or we can say it just how biology is.[1]

Although this paper is about explanation, it is not about the difference between explanation and interpretation in biology. When I use one of these words in this paper, I may sometimes mean both together. I think almost every explanation has some amount of interpretation, because I believe there are usually some values in explanations. I think value-free science is impossible and values may interfere

[1] For some information about pluralism and sciences, see "Chapter 1: The Miracle of Monism" in Dupré, 2012.

with scientific practice at many steps in the process, so I give great importance to having good values in science (respecting human rights, all living things, and the environment).

The main aim of this paper is to give a clear ground of the causation and explanation in plant phenotype research.[2] I argue that its grounding properties are based on two aspects: (1) the complex interactions of processes in phenome[3] ontology, and (2) the complex structures of society and the scientific community. Although these are not new in philosophy of biology, the novelty of my contribution is in examining them specifically in plant research, an area which has received much less attention in philosophy of biology literature than animal research.[4] This paper is also significant due to the recent rise in plant phenomics. The drastic changes in environment (i.e. climate change, fast population growth) and the related need for more hardy, nutritious, and voluminous crop plants has generated new challenges for plant scientists and spurred advancements in phenomics. After giving a detailed explanation of phenome occurrence and research (the second part), the final part of the paper compares current phenotype research and Aristotle's Theory of Four Causes. I believe this attempt is helpful for constructing a stronger ground on the subject, since Aristotle is one of the earliest philosophers of biology, and created very well-organized principles for the scientific study of life (Lennox, 2006). His theory of investigation and explanation of living things still influences many philosophy of biology theories,[5] some of which I refer to in the last part. Although there are many discrepancies between the Four Causes theory and causation and explanation in current phenotype research, I found this comparison useful since these well-organized principles on investigating living things have a structure that emphasizes both multiplicity[6] and strong relatedness of causes.

2. Phenome and Phenome Research

A phenome is the expression of an organism's self in a certain environment; it is everything of the organism except its genome. The important thing to remember about an organism and its environment is that we usually cannot define a clear boundary between them; they are interpenetrating, constantly sending signals back and forth, constantly causing and affecting one another. It is usually very difficult to define the "self" of an individual organism because its environment is so intrinsic to it that we cannot easily say where the organism ends and its environment begins.[7] For example: arbuscular mycorrhiza and plants, or microbiota in animals' intestine. Both the arbuscular mycorrhiza and the microbiota of intestines have different genotypes than the plant's genotype and the animal's genotype, but they are part of the organisms' systems, and every phenotypic trait of the organism has a direct or indirect effect on them. We can even say that a phenotypic trait—one part of the phenome that we decide to measure—of an animal is also a phenotypic trait/environment of its microbiota.

Before proceeding to the occurrence of the phenomes and causation and explanation in phenotype research, I would like to give a few more definitions of "phenotype" for the sake of clarity: "The phenotype is the descriptor of the phenome, the manifest physical properties of the organism, its physiology, morphology and behavior" (Lewontin 2011). "The appearance or characteristics of an organism resulting from both genetic and environmental influences" (Nicotra et al. 2010). A phenome is "The expression of the genome as traits in a given environment" (Furbank and Tester 2011). Lewontin says that phenome to phenotype is like genome to genotype and token to type (Lewontin 2011). For example, having brown hair is a phenotype, but my brown hair is my phenome.

A phenome occurs through the interaction between a plant and its environment. Plants constantly

[2] While this paper is about plant phenotype research, the thesis I am asserting here is coherent with causation and explanation of phenotype research into all the living things.

[3] I would like to acknowledge that my ideas on "the plant phenome" have been developed through my years of research experience in plant biology and physiology laboratories with several groups of colleagues and through reading a great number of biology papers, not all of which I can refer to here. Also, my recent meetings with John Dupré helped me very much in framing my ideas.

[4] Like plants, working on microbes is also a new trend in philosophy of biology.

[5] For instance: Mayr 1961; Short 2002; Pigliucci 2003; Lennox 2006; Anzaldo 2007; Williams 2010; Haig 2014; Mix 2015.

[6] A multiplicity of causes—via mentioning Aristotle—have been pronounced recently, for instance: proximate and ultimate causes in biology (Mayr 1961) and Tinbergen's four questions in ethology: function, phylogeny, mechanism, ontogeny (Tinbergen 1963).

[7] There is some recent literature on the self of the living things. For instance: Pradeu and Vitanza (2012).

receive signals from their environment. These signals are classified according to their physical and chemical properties and quantities, and converted from one form to another and transmitted through the plant's body. During these transmissions, many internal processes are rearranged such that each process affects another (influences one or more other processes) from some specific parts, resulting in changes to these parts and causing ramified cascades. These rearrangements of internal processes are *a response to environment*, as plant scientists call it. *This response is basically the plant itself*. It can be many things, such as opening or closing stomata, doing more or less photosynthetic activity, expressing more or less of a gene, producing more or less of a hormone, etc.

Just like plants, the environment senses signals, too. There is again a signal transmission and response production, but these happen differently in living and non-living parts of the environment. So there is a constant relation: reception → response production → reception of the produced response → again response production → reception → production → reception → production… between a plant and its environment. This interaction can be illustrated though an arrow pointing two ways between plant and environment, but if we want to better illustrate it, we would draw *many* two-way arrows (we do not know for sure how many), and these arrows would usually touch each other as well, constituting a very complex net. The important point is that this net of interaction is not just between plant and its environment; it is also inside both of them. That is also why there is no very clear boundary between a plant and its environment.

Plants (and also their environments) have many parts and levels—with every level having many parts, and every part having many levels—which are constantly affecting each other. Both plants and the environment have complex pathways of interaction in and between every level and part of them. Because of this complexity, in the occurrence of phenotypes there are sets of causes which are interrelated. Therefore, phenotype explanation and research is clearly *context dependent*. Scientists look at this complexity and then decide to research a *single* response of a *single* organism to a *single* change in a *single* environmental parameter or a *single* genome part. So scientists have a clear *purpose* because they carefully define their research parameters. This purpose is also context dependent: it depends on the state of scientific knowledge and the state of society (these are also complex systems—even more complex than phenotype occurrence). When a scientist defines these parameters and designs an experiment, it means that she/he picked a *possible cause* from the complexity of the phenomena. Of course she/he does not pick it randomly; this decision depends on scientific knowledge and society. Then she/he designs an experiment with several groups in which all the values are stable—"natural" conditions (background conditions)—and she/he makes an *intervention* (Woodward 2010) with her/his nominated possible cause in some of the groups. At the end of the experiment, if her/his hypothesis comes up as not false (she/he observes a change in the subject phenotypic trait in different groups), she/he concludes that the possible cause is actually the cause of the subject phenotypic trait. Two very important points here: (1) this is a cause of the phenotypic trait in question in *a specific context* (in her/his experiment context), and (2) she/he *nominated it as a cause* in the first place; so she/he should be aware that there are many other possible causes that may be more important than the subject cause. This is why there is always a very detailed materials-methods section in biology research papers. The specific context of their experiments is given both in resulting explanations (in general terms) and also in materials-methods section (in detail).

Before giving a simple example, I would like to say that from different areas of phenotype research, many very different examples can be given.

For instance, say there is a plant physiology research group that wants to investigate plant responses to rising carbon dioxide (CO_2) levels in the atmosphere. And suppose that the following describes the situation of the society and the scientific community:

- We have climate change problem, and we now know that it is because of greenhouse gases emissions from human activities.
- We have climate change research and action groups.
- CO_2 is one of the greenhouse gases (current concentration is around 400 ppm[8] in the atmosphere).
- We hope to stop or slow climate change (to reduce our greenhouse gases emissions). We also want to be prepared for possible scenarios: until the end of this century, in the best-case scenario the CO_2 concentration will be

[8] pm: parts per million

between 430-480 ppm, and in the worst-case scenario it will be more than 1,000 ppm (IPCC 2014).

- We have a huge food security problem: many people do not have enough food, and many more are suffering from nutrient deficiency related diseases.
- Wheat is one of the main crop plants.
- Plants use CO_2 in photosynthesis. (More CO_2 may be beneficial for production?)

There is a clear need in society for better-adapted and more nutritious crop plants. So this creates a challenge for plant science community[9]: to more thoroughly understand how plants cope with changes in climate (such as elevated levels of CO_2). Let's say the example lab group designs this experiment: research on a durum wheat cultivar's *Triticum durum*, Sarıçanak 98 *(certain organism's)* grain yield, nutrient composition in grain, photosynthesis activity,[10] etc. *(certain responses—certain phenotypic traits)* in response to 500 ppm and 800 ppm *(certain change)* CO_2 concentration *(certain environmental parameter).* They will have three groups: 400 (as current ambient—control group), 500 and 800 ppm (as elevated: possible future concentrations). They will try to hold all the other parameters (parameters that we know: temperature, humidity, soil conditions, light, etc.) at optimum stable values,[11] and measure the traits in question in these three groups. They will use as many replicates as they can in each group so that their results will be more reliable. At the end of the experiment, if they find statistically significant differences between the measured values from each group, they will calculate the probabilities and conclude by explaining how this certain change (100 ppm and 400 ppm difference from current conditions) in this certain environmental parameter (CO_2 concentration) causes a certain phenotypic trait (let's say: bigger grains—and since they are investigating several related phenotypic traits, they may explain some mechanisms) in this durum wheat cultivar (*Triticum durum* Sarıçanak 98). They are totally aware that this result is for *this specific experiment context*.[12]

If they were to extend their research, they may want to get closer to the "real conditions," so they may want to experiment on a combination of changes. According to current scientific knowledge, climate change is causing a combination of changes in the environment. For example, in some parts of the world fields are facing drought and high temperatures, while in some other parts they are faced with floods, etc. So let us suppose that this example research group decided to work on temperature stress as well. When they design an experiment with both elevated CO_2 concentration and high temperature, they will have many more experiment groups,[13] and with replicates in each group they will end up with a very big experiment and a lot of data.[14] This time their explanation will be broader, but still within their experiment context.

3. Aristotle's Theory of Four Causes and Phenome Investigation

In *Physics* book 2, Aristotle begins by stating the difference between things that are natural and things that are not natural (Aristotle, *Physics* 192 b_{8-16}). He emphasizes that natural things have a "*source of change and of stability*" inside them. Later he makes this emphasis clearer by saying:

> The nature of a thing, then, is a certain principle and cause of change and stability in the thing. (Aristotle, *Physics* 192 b_{20-21})

He further says that if we want to have the knowledge of a thing's nature, we must investigate it:

> For the point of our investigation is to acquire knowledge, and a prerequisite for knowing anything is understanding *why* it is as it is – in

[9] Of course the relation between science and society is much more complicated than in this simple example.

[10] As stated before, the decision to choose these phenotypic traits depends on the current state of scientific knowledge and society; in this case, these are agriculturally valuable traits.

[11] This is usually not completely attainable because of the complexity of phenotype occurrence, so it can be done only to some extent. Because of this, resulting data will probably contain a lot of "noise."

[12] Even if only one factor was different in the background conditions, the results could be different.

[13] Let us say that they add two elevated temperature groups (control: 24 °C, elevated: 25 and 27 °C) to the previous design. They would then have the following groups: 24 °C–400 ppm CO_2 concentration; 24–500; 24–800; 25–400; 25–500; 25–800; 27–400; 27–500; 27–800. If they were to add one more parameter to the experiment, say drought stress, they would have many more groups, also in combination with three of the stressors.

[14] They will analyse this data and they will try to explain both sole effects and the interactive effects of the subject possible causes on the subject phenotypic traits.

> other words, grasping its primary cause. (Aristotle, *Physics* 194 $b_{18\text{-}20}$)

Then he tells us that there are four ways of using the word *cause*, and these are:

- Material Cause: "from which a thing is made and continues to be made" (Aristotle, *Physics* 194 $b_{23\text{-}24}$). He says that we ask the question "What is it made of?" to investigate the material cause of a thing.
- Formal Cause: "for the form or pattern (i.e. the formula for what a thing is" (Aristotle, *Physics* 194 $b_{26\text{-}27}$), and we ask "What is it?" to investigate the formal cause of a thing. This also gives us the definition of a thing.
- Efficient Cause: "for the original source of change or rest." (Aristotle, *Physics* 194 b_{29}), and we ask "What initiated the change?" to investigate the efficient cause of a thing.
- Final Cause: "for the end. This is what something is for" (Aristotle, *Physics* 194 b_{32}), and we ask "What is it for?" to investigate the final cause of a thing.

He further says:

> So it is clear that there are these causes and there are this many of them. It is the job of the natural scientist, then, to understand all four of these causes; if he refers the question 'Why?' to this set of four causes—matter, form, source of change, purpose—he will be explaining things in the way a natural scientist should. (Aristotle, *Physics* 198 $a_{21\text{-}23}$) (also referred to in Falcon 2012)

I think that when we read the *Physics* we see that there are no *clear boundaries* between causes, that they are very closely related to each other; they sometimes may refer to the same cause (for example: the formal cause may be the final cause), but yet there are also *some* differences, and they are related to each other in different ways. In one part of the book, Aristotle says that formal cause, final cause, and efficient cause are same:

> In many cases, the last three of these causes come to the same thing. What a thing is and its purpose are the same, and the original source of change is, in terms of form, the same as these two. (Aristotle, *Physics* 198 $a_{24\text{-}26}$)

He also points out the difference between material cause and formal cause while again stating the sameness of formal—definition—and final—end—causes. He says:

> …the necessity is in the matter, but the end is in the definition. (Aristotle, *Physics* 200 a_{14})

He further says that sometimes having knowledge of three of the causes (or even just one) is enough to have knowledge of a thing, so that if we know a thing's material cause, formal cause, and efficient cause, we have the knowledge of it:

> In short, then, the question 'Why?' is resolved by answering it in terms of a thing's matter, what it is and its original source of change. (Aristotle, *Physics* 198 $a_{31\text{-}32}$)

A living thing consists of levels of materials, and without these special materials and their specific features, there cannot be that specific phenotype. Because of this, an explanation of or research into any level or part of an organism (molecules, atoms, subatomic particles, organs, tissues, etc.) is similar to Aristotle's explanation of material cause. Aristotle says: "necessity is in the matter." Today we know necessity is in the matter together with interactions/encounters. We know that each material, for example each atom, has some specific features that necessitate a specific move depending on the moves/features of the neighboring atoms, and the combination of some specific atoms constitutes a molecule (another level) that has some specific features necessitated from its specific parts (features of the atoms that constitute it)[15] and the moves/features of neighboring molecules. In other words: every part of an organism has its specific roles/features according to its features and its surroundings/position within this net, so that if it was in a different place, its roles/features/activities would have been very different. Since any genome, phenome, or environment consists of many kinds of materials (nucleic acids, phosphate groups, sugars, proteins, tissues, soil particles, water, etc.), research on them and explanation of them resembles Aristotle's investigations and explanations of material causes.

Genes carry information that can be transferred between generations, and genotype sometimes defines the organism's species, so we can say: research into and explanation of genotypes is similar to investigations and explanations of formal causes. But I want to emphasize that knowing the genotype of an

[15] The features of the molecule depend on the features of its atoms, but is different than the combination of the features of its atoms.

organism does not mean that we know its phenotype. It means that we have knowledge of one of the causes of the organism's phenotype. Through their phenome, genes interact with the environment to express themselves, and how much and which gene products will be produced depends on environmental signals and the developmental stage. So we can say that research into and explanation of environmental factors is also similar to investigations and explanations of formal causes. But as with genotypes, we cannot say that we have knowledge of the subject phenotype when we only know about environmental factors that affect it. Again we can only say that we "have knowledge of *some* of the causes of the subject phenotype." Phenotype sometimes defines the organism's species and its morphology and appearance, so research into and explanation of phenotypes is also similar to investigations and explanations of formal causes.

There cannot be a phenome without environmental effects. There cannot be organism without an environment, and I think we can even say that environment is intrinsic to living things. The interaction between organism and environment causes changes in living things, so research into and explanation of environmental factors and organisms (phenome+genome) is similar to investigations and explanations of efficient causes.

If we want to compare Aristotle's Theory of Four Causes and current phenotype research, we can say that Aristotle's material, formal, and efficient causes are similar to causes in phenotype research in investigations of phenotype, genotype, and environment (and in investigations of all related parameters: ecological, evolutionary, developmental, molecular, genetic, epigenetic, physiological).[16]

In phenotype occurrence there is both *stability and change* in organisms. Aristotle says these are both in the nature of the natural things. I think this togetherness of stability and change in living things occurs through the interaction between an organism and its environment, and that these are the very basic and essential features of living things. As parts of living things, genotype, phenotype, and environment have both stability and change in them. Speaking very generally, genotype has more stability, while environment has more change, and phenotype has equal parts of both.

Shields (2008) points out that "*some scholars have come to understand four aitia more as explanations rather than as causes*," and also says: "*Aristotle's approach to aitia may be regarded as blurring the canons of causation and explanation*." I think causation and explanation are strongly related to each other, but they are different in the sense that causation is about "what is there" and "what really cause something," while explanation carries both causation and our *intentions* (although causation also carries our intentions in the sense that we choose to give our attention to some specific causal relations rather than others).[17] Still, the causal relation is really there in the phenomena we are investigating.

Because of this, if we compare Aristotle's Theory of Four Causes and current phenotype research, one of the four causes—final cause—has no analogue in contemporary science, since we do not know (and *cannot* know) if it really exists. According to contemporary science, there is no *telos* in biological phenomena. Living things and their parts are not moving toward some specific end. They just happen because of the situation (all the current interactions) and the history of the situation (evolution and the all past interactions). Aristotle says that there is no evolution; he believed that each living thing has a specific form which is also its *telos* (final cause).

Today we know that this is not true. However, in our investigations into the living things we do use functional language: we say that some parts of living things are *for* some other parts or activities, as if they have some specific *telos*. But we speak this way due to the limitations of language, because this is how we inquire into living things (it is in our *explanations*). In attempting to understand how a living system works, we try to systemize a set of norms which operate it. However, we sometimes observe that it does not follow certain norms.[18] For example, if a living thing encounters a very unusual environmental stimuli, it may not act as we expect; it will

[16] Pigliucci (2003) says evolutionary processes are similar to efficient causes. This is coherent with what I am saying, but I think they are not the only processes that are similar to efficient causes. As I have stated, the investigation of the current interaction and the history of interaction between an organism and its environment resembles investigation of efficient causes. Pigliucci also states that ecological niche construction resembles final causation. I may agree with this only in the explanatory sense, not as a causal relation.

[17] And again I would emphasize that in the case of a regular phenome investigation, we leave many other parameters aside and attend to only a few. It is our *choice* to investigate *certain causal relations* rather than others.

[18] This is related to "reasons of indeterminacy in biology" (Mayr, 1961).

interact with the stimuli (probably stretching its phenome),[19] and through this interaction (which happens within the context of many other interacting things) it will rearrange itself and express itself according to this new environment, so its phenome will change. Alternatively, it may simply die (although when a system dies, it turns into another system). It does not have a specific end that it tends to be; it may become a vast number of things—different phenomes—depending on its interactions.

Living things are obviously different than non-living things, but this does not mean that they have a *telos*. I think the only things that living things possess that resemble a *telos* are the needs *to exist* and *to express themselves*. This expression is always in some specific situation, and depends on all the organism's interactions and the history of all interactions. We cannot say that there is a *specific self* or *end* that living things tend to be. Some philosophers argue that biological organisation is inherently teleological, and that this intrinsic teleology can be grounded in the process of evolution; they say the biological organisation determines itself (Mossio and Bich 2014). I argue that it does not determine itself; rather, it expresses itself in its environment through its interactions. The self is the phenome of the biological organisation at a certain specific point in time (or the collection of the phenomes of it through specific time periods), and its expression is the result of the biological organisation-environment interactions (both the current interaction and the history of all interactions: a biological organization possesses the effects of its ancestral history throughout its genome and epigenome, and also its own history throughout its epigenome and phenome).

As stated above, there is an analogue to final cause in phenotype explanation and research (not in occurrence and not in causation), and we can also say that a *scientist's purpose* resembles a final cause. But the degree of similarity between final cause and a scientist's purpose is much less than the degree of similarity between formal-material-efficient causes and phenotype-genotype-environment parameters. This "final cause analogue" cannot be (and should not be) treated in the same way as the "formal cause analogue" in phenotype research, as are final and formal causes of in Aristotle's Theory, because if they are same, then the scientist's purpose is interfering with the explanation of the experiment.

The gene-centered view can be given as an example: here the scientific value of the experiment's data is still good (a causal relation is revealed), but the scientist's purpose is interfering with the correctness of the explanation. This happens when a scientist sees her/his nominated cause as the most important cause of the subject phenotype, as if there is no complex net, no set of many causes in the occurrence of phenotype. In this case, the scientist's nominated cause becomes like Aristotle's final cause in suggesting that the gene *has a purpose*: to *cause or try to cause a certain phenotype*. We know that genes do not have purposes, and they do not strive toward specific ends; they merely interact with a phenome and environment in a way that causes specific proteins to be produced. I am not saying that in no case is a gene the most important cause of the subject phenotypic trait. In rare cases a gene may be the most important cause,[20] but there are also other causes that altogether constitute the specific situation in which subject phenotype occurs. And in these rare cases in which a gene is the most important cause of the subject phenotype, scientists must investigate many factors to reach this conclusion.

Both "Aristotle's four causes" and "genotype-phenotype-environment" or "genetic, epigenetic, physiological, developmental, molecular, environmental, evolutionary factors" are related to each other very strongly in that they describe a complex net of factors that altogether constitute the organism. We can look for specific causes in specific research plans, depending on the situation and/or the purpose of the investigation, but we should always be aware that this is a way of doing practical research, and that these specific causes may differently affect other contexts, or there might be some other causes which are more important than the subject cause. If our aim is explaining an organism's phenotype as a whole (all the causes of it),[21] we should search for all three causes: formal-material-efficient (or ecological, physiological, developmental, molecular, epigenetic, genetic, and evolutionary) factors in the context of their interrelated state. And whether we are researching one part of the phenotype, or we are researching all parts of it, an awareness of all the interrelated factors and the complexity of the phenotype is very important for attaining a good explana-

[19] This stretching ability (phenotypic plasticity) also depends on all the previous encounters—all interactions the living system had throughout its life, and all interactions of its ancestors.

[20] Maybe we can say that in these cases, genes are "the difference makers" (Waters, 2007).

[21] This is of course still context-dependent.

tion of the phenotype, just as Aristotle's student of nature is aware of the multicausality of nature.

References

Anzaldo A A (2007) Back to the Future: Aristotle and Molecular Biology. *Ludus Vitalis* Vol XV, 28: 195–198.

Aristotle (2008) *Physics*. Oxford World's Classics, Translation: Robin Waterfield.

Dupré J (2012) *Processes of Life*. Oxford: Oxford University Press.

Falcon A (2012) Aristotle on Causality. In *The Stanford Encyclopedia of Philosophy,* ed. Edward N. Zatla.

Furbank R T, Tester M (2011) Phenomics – technologies to relieve the phenotyping bottleneck. *Trends in Plant Science* 16,12: 635–644.

Haig D (2014) Fighting the good cause: meaning, purpose, difference, and choice. *Biol Philos* 29:675–697.

IPCC (2014) Climate Change, Synthesis Report of the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. (eds.) The Core Writing Team, Rajendra K. Pachauri, Leo Meyer (subject to final copy edit and layout).

Lennox J (2006) Aristotle's Biology. In *The Stanford Encyclopedia of Philosophy,* ed. Edward N. Zatla.

Lewontin R (2011) The Genotype/Phenotype Distinction. In The Stanford Encyclopedia of Philosophy, ed. Edward N. Zatla.

Mayr E (1961) Cause and Effect in Biology. *Science* 134:1501–1506.

Mix L (2015) Nested explanation in Aristotle and Mayr. *Synthese* July: 1–16.

Mossio M, Bich L (2014) What makes biological organization teleological. *Synthese* November: 1–26.

Nicotra A B, Atkin O K, Bonser S P, Davidson A M, Finnegan E J, Mathesius U, Poot P, Purugganan M D, Richards C L, Valladares F, van Kleunen M (2010) Plant phenotypic plasticity in a changing climate. *Trends in Plant Science* 15: 684–692.

Pigliucci M (2003) From molecules to phenotypes? The promise and limits of integrative biology. *Basic and Applied Ecology* 4: 297–306.

Pradeu T, Vitanza E (2012) *The Limits of the Self: Immunology and Biological Identity*. Oxford: Oxford University Press.

Shields C (2008) *Aristotle. The Stanford Encyclopedia of Philosophy* (First published 2008; Summer 2012 Edition), Ed.: Edward N. Zatla.

Short T L (2002) Darwin's concept of final cause: neither new nor trivial. *Biology and Philosophy* 17:323–340.

Tinbergen N (1963) On aims and methods of ethology. *Zeitschrift für Tierpsychologie* 20:410–433.

Waters C K (2007) Causes That Make a Difference. *The Journal of Philosophy* 104: 551–579.

Williams N (2010) Aristotle's Lagoon. *Current Biology* 20(3): R84.

Woodward J (2010) Causation in biology: stability, specificity, and the choice of levels of explanation. *Biol Philos* 25:287–318.

THE PRINCIPLE OF COMMON CAUSE AND ITS ADVANTAGES AND LIMITATIONS IN SCREENING THE CORRELATED EVENTS OFF

Joby Varghese
Department of Philosophy
School of Humanities
University of Hyderabad
Gachibowli, Hyderabad, Telangana
India - Pin- 500046
e-mail: jobyv@uohyd.ac.in

Abstract:
The Principle of Common Cause (PCC) puts forward the idea that events which occur simultaneously and are correlated have a prior common cause which screens off the correlation. I endorse the view that the PCC does qualify as a principle that can be used as a tool in explaining improbable coincidences. However, though there are epistemological advantages in common cause explanations of correlated events, the PCC is not impeccable. This paper offers a preliminary assessment of the PCC advocated by Reichenbach, and then attempts to illustrate three scenarios in which the principle might be inadequate in explaining correlated events. The paper also compares the Common Cause Principle and the Causal Markov Condition (CMC), and examines the advantages of CMC over the Common Cause Principle.

Key words: Common Cause Principle, screening off, correlated events, limitations of the PCC, Causal Markov Condition.

1. Introduction

In the history of philosophy of science, theories of explanation have often been associated with causation. That is to say, to explain a phenomenon or an event is to identify its cause. Several philosophers of science think of explaining events in relation to causation. In other words, any phenomenon or event might be adequately explicated if one is able to identify its cause. Thus, for long years the concept of causality has played an indispensable role in explaining events. The advocates of this view are Salmon (1980, 1984), Humphreys (1982, 1989), Lewis (1986), Kitcher (1989), Woodward (1999, 2003), etc. There have been many suggestions about how we must relate causes and effects to provide causal explanations.

A crucial factor which plays a significant role in causal theories of explanation is the principle of common cause (Reichenbach 1956). The Principle of Common Cause (PCC) plays a vital role in explaining a phenomenon in terms of it causes. We seek a common cause when we are confronted with instances which appear to be improbable coincidences. The PCC is important because invoking the help of a common cause will enable us to screen off one correlated effect from the other. In every such case, the thing which does this screening off will always be a common cause of two or more joint effects, not a common effect of joint causes. We must be careful when we are in search of common causes because there can be cases in which we might mistake a common cause for something else, and that might, in turn, produce adverse repercussions on our attempt to screen off effects by a common cause.

This paper critically examines the common cause principle and gives a detailed account of the it, along with the notion of "screening off." The paper focuses on some of the shortcomings of the PCC, and analyzes three scenarios in which the employment of "screening off" via the PCC is prone to happen, and that might, in turn, lead to incorrect explanations. This paper also compares the PCC and the Causal Markov Condition (Hausmen and Woodward 1999; Spirtes, Glymour and Scheines 2000; Pearl 2000), and examines its advantages over the common cause principle.

2. Reichenbach's Principle of Common Cause

The first glimpses of the notion of common cause can be traced back to the work of Bertrand Russell (1948). According to him, when there is a group of complex events in more or less the same neighborhood, and independent events having a common structure occur, it is probable that they have a common causal ancestor. He elaborates with an example: in different parts of the country, a number of middle-aged ladies mysteriously die in baths after getting married and purchasing life insurance. An investigation based on the identical structure of the incidents leads to the supposition of a common causal origin, Mr. Smith, who is later convicted and hanged.

The Principle of Common Cause was formally developed and structured by Paul Reichenbach. The principle states that whenever there occurs an improbable coincidence, there is a common cause (Reichebach 1956). According to the PCC, any correlation between two factors which do not seem to be directly influencing one another is due to their (possibly hidden) common cause. Note that this is a purely probabilistic formulation. An instance of such an improbable coincidence is as follows: when two lights in the same room go out suddenly, an immediate reaction is to look out for a burned fuse or some kind of interruption of the power supply. Reichenbach thus argues that the improbable coincidence is explained as the product of a common cause. The principle is illustrated as follows:

> For simultaneous events A and B, $Pr(A\&B) > Pr(A) \times Pr(B)$, there exists an earlier common cause C of A and B, such that $Pr(A/C) > Pr(A/{\sim}C)$, $Pr(B/C) > Pr(B/{\sim}C)$, $Pr(A\&B/C) = Pr(A/C) \times Pr(B/C)$, and $Pr(A\&B/{\sim}C) = Pr(A/{\sim}C) \times Pr(B/{\sim}C)$[1] .

Here, C is said to "screen off" the correlation between A and B when A and B are uncorrelated conditional upon C. So, the principle states that events which occur simultaneously and are correlated have a prior common cause which screens off the correlation. In an informal way, we can say that there may be events occurring simultaneously, and there might be a surprising correlation between them; then this striking correlation is assumed to be a result of their common cause, which is sometimes hidden. When there are correlated events which are known as improbable coincidences, those correlated events are explained with reference to a common cause. Whenever we see two events frequently and simultaneously occur, often the reason is that one of them causes the other, or a third thing causes both, i.e., most apparently improbable coincidences are produced by a common cause. In general, when we find a correlation between two events, A and B, we might not take into consideration the direct causal connections between those events because there exists a common cause due to which the correlation occurred. For example, a chain of bomb explosion takes place in a city, and the police infer a common cause behind this incident. Here, the presence of common cause can be invoked in order to explain such improbable events. Similarly, when there are such kinds of correlations among the events, we construct an explanation that poses a common cause, or an explanation that would posit separate causes for the incidents. If we glanced at both explanations, we would find that the common cause explanation may be superior in two ways. It may be more probable, and it may be more explanatory.[2] Hence, whenever an explanation is given in terms of a common cause, what we essentially do is formulate an epistemological assumption about the conditions under which a particular kind of causation can be non-deductively inferred.

3. Screening off: A Unique Probabilistic Relationship

An important element that goes along with the Principle of Common Cause is the notion of "screening off." The terminology "screening off" describes a particular type of probabilistic relationship.

If $Pr(E \mid A \& C) = Pr(E \mid C)$, then C is said to screen A off from E. Reichenbach used the apparatus of screening off to address the problem of spurious correlations. That is, if event A and event E are spuriously correlated, then event A will be screened off from event E by a common cause.[3] Wesley

[1] Paul Reichenbach, 1956 pp. 158–159.

[2] Elliott Sober, 1984, p. 212.

[3] The example given by Reichenbach goes like this. A drop in atmospheric pressure (C) causes both a drop in the level of mercury in a barometer (A) and a storm (E). The atmospheric pressure will screen off the barometer reading from the weather: given that that the atmospheric pressure has dropped, the reading of the barometer makes

Salmon (1984) construed the relation of a common cause "C" to its joint effects "A1" and "A2" in terms of the idea of a conjunctive fork. According to Salmon's theory, the common cause must screen off each effect from the other; i.e.

Pr (A1/C & A2) = Pr (A1/C)

Pr (A2/C & A1) = Pr (A2/C).

The idea is that a common cause must render the joint effects probabilistically independent. Most importantly, the thing which does the screening off occurs before the correlated events. Thus, assuming that the screener off is the cause rather than the effect, the Common Cause Principle tells that the cause occurs before the effects. It says that there is a common cause "C" which gives rise to two entirely independent effects "E1" and "E2."[4]

I will illustrate this idea with an example. The students of a particular class are informed that their music teacher—who all the students in the class love—is absent on a given day. Having heard the news, all the students in the class become sad. Since both you and I are in the same class, we also become sad by the absence of the teacher. There is a correlation between my sadness (E1) and yours (E2). We cannot judge it to be merely a coincidence, because there is a common cause behind it which is "the absence of the music teacher" (C). We may also argue that the probability of you being sad is the same whether or not I am sad, given this situation in which the music teacher is absent. Hence, it is clear that the common cause (*the absence of the music teacher*) screens off one correlated effect (*E1* and *E2*) from the other. Hence, the notion of screening off plays a vital role in providing explanation by way of the Principle of Common Cause, according to which there must be a screening-off common cause for every set of correlated joint effects.

Bas Van Fraassen (1982) elaborates the way in which the common cause explains away its joint effects via screening off. Consider the correlation between heavy smoking and lung cancer. The correlation found in this instance is the simultaneous presence of two factors, which are *being a heavy smoker* and *having lung cancer*. A common cause, in this case, is a history of smoking (C) which produced both an addiction to smoking (E1) and unhealthy lungs (E2). A significant feature of a common cause is that the effects which stem from the common cause are independent. Hence, present smoking is an indication of lung cancer among the people as a whole. In a nutshell, the explanatory power of a common cause is as follows: (a) the common cause screens off the correlation between the independent effects (E1) and (E2), (b) the common cause has a positive statistical relevancy to both (E1) and (E2), and (c) the common cause entails a positive correlation between (E1) and (E2).

4. Problems with the Common Cause Principle

Though the Principle of Common Cause appears to be an adequate principle in explaining the occurrence of correlated events, there are some problems with it. The following section will critically scrutinize three scenarios in which the PCC might fall short.

4.1 Scenario I

A serious concern with the PCC is its viability in explaining correlated events when there is an *irrelevant correlation* between those events. This may happen when there is a common cause which produces an effect, and that effect then appears to be a common cause for another event. Though this effect appears to be the common cause, we may screen off this apparent common cause as an irrelevant one in relation to the real common cause. For example, in a rural area in which the villagers themselves produce their own food, suddenly there is widespread starvation. The people find the common cause behind this instance to be the collapsed economy. Then there comes a group of social workers who conduct a study about the situation in the village and of the people. Eventually, they reach the conclusion that the real cause of the starvation (S) is not the collapsed economy (E), but a rapid growth of the population (G) in the last few years, which caused a famine. This situation can be depicted like this:

Pr (S/E&G) = Pr (S/G)

Hence, the sudden growth rate of the population is the common cause of both the collapsed economy and starvation in that particular village. In this case, the correlation between the collapsed economy and

no difference for the probability of whether a storm will occur. A drop in the column of mercury (A) raises the probability of a storm (E). Actually, it does not raise the probability of a storm when we take atmospheric pressure into account. That is, if A and E are spuriously correlated, then A will be screened off from E by a common cause.

[4] Frank Arntzenius, 1992, p. 227.

famine turns out to be irrelevant, because the collapsed economy itself is caused by the rapid growth rate of the population, which may be screened off. Elliott Sober (1988) has also raised a similar concern.[5] He argues that similar laws of evolution of otherwise independent quantities can lead to correlations for which no common cause exists.

4.2 Scenario II[6]

Another issue the Principle of Common Cause may encounter and fail to address is when a single event gives the impression that it is more than one event occurring simultaneously, and thereby correlated when it is actually not. Such an instance can be illustrated as follows: there is a spinner, and the dial of the spinner is green in color, and the dial of the spinner is divided into three parts labeled with the even numbers 2, 4, and 6. The hand of the spinner moves, and after regular intervals, the hand stops. If we examine the moves and the stops of the spinner, apparently there are two events, occurring simultaneously;

Event A: The hand stops on green.

Event B: The hand stops on an even number.

The problem here is the overlapping of the events which are not really different events in the actual state of affairs. In this case, event A and event B are not *distinct due to an overlap* because whenever event A takes place, event B also simultaneously takes place, and there can be no cause that screens them off.

[5] Sober (1988) gives an example of positively correlated quantities. He suggests that similar laws of evolution of otherwise independent quantities can lead to correlations for which no common cause exists. His example goes as follows: the bread prices in Britain have been going up steadily over the last few centuries. The water levels in Venice have been going up steadily over the last few centuries. There is therefore a correlation between (simultaneous) bread prices in Britain and sea levels in Venice. However, there is presumably no direct causation involved, nor a common cause.

[6] Scenario II is a modified version of the criticism articulated by Lewis (1986). He argues that the PCC may fail if A and B are logically related, or involve spatiotemporal overlap. Suppose that events E, F, and G are completely unrelated, and that they are probabilistically independent. Define A to be E & F, and B to be E & G. Now A and B will be correlated, due to the logical overlap, and there may be no cause that screens them off.

4.3 Scenario III

It is quite probable that two more events might occur simultaneously and are correlated, and these events might be the effects of more than one ancestral or common cause. In such cases, the Principle of Common Cause might fall short in screening off the correlated events, because if two events A and B have more than one common cause, then there would be no one common cause that screens A off from B. This can be depicted by way of an example.

Suppose that there is a region where there is only one bread factory which supplies bread to the nearby localities. Unexpectedly, there is an increase in the price of bread (E1), and at the same time, the bread company faces loss (E2). This is a bizarre situation because generally the increase in the price of bread would bring more profit to the company, but instead the company faces loss. Now, a team is assigned by the company in order to conduct an inquiry to find out the causes behind the events (E1) and (E2). The investigation team does its work, then reports that there are two common causes behind the events (E1) and (E2) which occurred simultaneously and are correlated. The report shows that the factory workers were not motivated enough and produced less product than in previous years, and at the same time the workers demanded a salary increase which was eventually approved by the company (C1). Here, (C1) can be said to be the first common cause because it resulted in an increase in the price of the bread (E1) and the loss of the company (E2). Along with (C1), the team puts forward another cause which also prompted the occurrence of (E1) and (E2), namely that there was a severe drought which caused a wheat shortage (C2). (C2) is also a common cause of both the events (E1) and (E2). The problem the principle of common cause confronts here in screening off the correlated events (E1) and (E2) is that since there is more than one common cause, the PCC would not be able to screen off the correlated events with a common cause. The reason behind the claim is that even if the PCC screens off (E1) and (E2) with C1, they would still be correlated due to C2.

5. The Causal Markov Condition (CMC): A Causal Model

Some serious works on causation in Philosophy, Statistics, and Cognitive Science have made increasing use of a particular type of a model known as the Causal Markov Condition (CMC). Pioneers of the

CMC are Judea Pearl, Peter Spirtes, Clark Glymour, and Richard Scheines. The Causal Markov Condition is a certain type of causal model. Before we get into a discussion of the Causal Markov Condition, let us quickly define what a causal model is. Causal models take account of contributions from various disciplines such as statistics, artificial intelligence, philosophy, econometrics, epidemiology, psychology, and other disciplines. A causal model is an abstract model that describes the causal mechanisms of a system, and it must express more than a correlation, because correlation does not imply causation. Bob Rehdner (2003) argues that the central claim of a causal-model theory is that people's knowledge of categories includes not only category features, but also the representation of causal mechanisms that link those features.

A causal model is comprised of a set V of variables, and it also includes two mathematical structures characterized over V. First, there is a directed graph G on V: a set of directed arrows, having the variables in V as their vertices. Second, there is a probability distribution P defined over propositions about the values of variables in V.

The relationships in the graph (G) are often described using the language of genealogy. An "ancestor" of X is a variable positioned on the left of X along a directed path, and a "descendant" of X is a variable which lies to the right of X along a directed path. If there is an arrow which has its tail at Y and head at X, then we say that Y is a "parent" of X. On the other hand, if the arrow is from X to Y, then the variable X is a "parent" of Y. The causal structure among the variables is represented on the graph over the variables in V. For instance, a causal influence of X on Y that is not mediated by any other variable in the set V is represented by an arrow from variable X to variable Y. Here, that means X is said to be a direct cause of Y relative to the variable set V.

Secondly, the probability quantity P is specified over the propositions of the form X = x, where X is a variable in the set V and x is a value in the range of X. Probability measure P is also specified over conjunctions of negations, and disjunctions of such propositions. Hence, the conditional probabilities over such propositions will be well-defined whenever the event that is doing the conditioning has positive probability.

According to Pearl (2000), the CMC is a mathematical theorem. Pearl says that macroscopic systems are deterministic. In practice, we do not have access to all of the causally relevant variables affecting a macroscopic system. Even so, the inclusion of enough variables in our model will satisfy the CMC, and the set of the analytic tools for studying the system will be robust. In Spirtes, Glymour, and Scheines (2000), the CMC has more the status of an empirical posit. They defend the CMC in two different ways. Firstly, empirically it seems that a large number of systems do in fact satisfy the CMC. Secondly, many of the methods which are usually used for detecting causal relationships tacitly presuppose the CMC. For example, the use of randomized trials assumes a particular case of the CMC.

Woodword and Hausman (1999) characterize the CMC as follows:

> "X is a variable in the variable set V. Descendants(X) is the set of all the descendants of X in the graph, i.e., all the effects of X in the set V. Parents(X) is the set of all the parents of X in the graph, i.e., the set of all the direct causes of X in the set V. Since Parents(X) is a subset of V, direct causes of X that are not in V will not be in Parents(X). Independence is, of course, probabilistic independence. So, according to the CMC, conditional on the set of all its direct causes in V, X is independent of all the variables in V that are not direct causes of X or effects of X."[7]

Further, conditional on its parents, X is independent of every variable in V except its effects. Thus every variable is screened off by the set of all its parents from every other variable except its effects, i.e., for all Y distinct from X, if X does not cause Y, then Pr(X/Parents (X)) = Pr(X/Parents(X) · Y).

5.1 Screening-off and the Causal Markov Condition

The CMC is a generalization of familiar screening off conditions. The CMC suggests that there is a full set of common causes which screens off joint or correlated effects from one another, which directly causes a screening off of effects from distal causes, and those variables which are not connected as cause and effect or as effects of a common cause are probabilistically independent. Woodward and Hausman (1999) also state that according to the CMC, if Parents (X)—i.e. the subset of V containing all the direct causes of X in V—is non-empty, then X is

[7] Daniel Hausman and James Woodward, 1999, p.523.

probabilistically independent of every variable except its effects, conditional on Parents (X). They also argue that Parents(X) screens off X from its indirect causes and from other effects of Parents (X) which are not also effects of X. For instance, if certain probabilistic dependence relations hold, then certain causal facts must also hold. In that case, the PCC is precisely of this form, which acknowledges that the PCC follows from the CMC. In fact, the Causal Markov Condition is a modified version of the Common Cause Principle. It goes as follows: if $Q_1$and Q_2 are correlated, and Q_1 is not a cause of Q_2, and Q_2 is not a cause of Q_1, then there are common causes of Q_1 and Q_2 in the set $\{Q_1,\ldots, Q_n\}$ such that Q_1 and Q_2 are independent conditional upon these common causes. Woodward and Hausman also argue:

> "The CMC puts forth a promising formulation of the intuition behind the claim that common causes and causal intermediaries screen off. It is false to say simply that common causes screen off their effects or that causal intermediaries screen off their effects from their causes. A common cause of X and Y or a causal intermediary between them obviously does not screen them off when there is some independent causal relationship between them. For example, a variable Z can fail to screen off X and Y even when the only causal connection between X and Y is via Z."[8]

Likewise, the CMC provides solutions to some of the difficulties with traditional formulations of claims about the notion of screening off put forward by Reichenbach (1956). However, even the solutions suggested by the CMC formulation might also appear to be unacceptable, because sometimes there can be problems regarding the assumption of causal sufficiency. For example, X and Y are not related as cause and effect, but they are effects of a common cause Z. Suppose that Z is not in the set of variables V, and that Z causes X via a causal pathway which does not navigate through any of the variables in V. In such a case, X and Y may well covary even though X does not cause Y. Here the CMC appears to be false. We can rule out this difficulty by confining ourselves to what Spirtes, Glymour, and Scheines (2000) call "causally sufficient" sets of variables, i.e., sets of variables that do not leave out relevant common causes. It is very much possible that in many practical applications, a causally sufficient set of variables will not be known, and we need a method for discovering causal relationships that does not rest on the assumption of causal sufficiency.

5.2 The Principle of Common Cause (PCC) and the Causal Markov Condition (CMC)

According to Reichenbach (1956), the common cause C is said to "screen off" the apparent correlation between events (E1) and (E2) when (E1) and (E2) are uncorrelated conditional upon C. Reichenbach's principle can also be formulated in the following way: when simultaneous and apparently correlated events occur, there is a prior common cause which screens off the correlation.

The Common Cause Principle is employed in cases in which the common causes and the effects are exhibited as sets of quantities with continuous or discrete sets of values, not single events which occur or do not occur. A simpler illustration is as follows: if the (values of) quantities (E1) and (E2) are correlated, then there can be any number of common causes C_1, C_2,…,C_n, such that conditional upon any combination of values of C_1, C_2,…,C_n at an earlier time, the values of (E1) and (E2) are probabilistically independent.

5.3 Advantages of the CMC over the PCC in Deterministic Systems

Imagine that the state of the world or a system of interest at any time determines the state of the world or that system at any later time. Arntzenius (1993) argues:

> "for any quantity X of that system at any time t, there will be at any other time t′, in particular, any later time t′, a quantity X′ such that the value of X′ at t′ uniquely determines the value of X at t. Conditional upon the value of X′ at t′, the value of X at t will be independent of any other value of any other quantity at any later time"[9].

Reichenbach's common cause principle thus fails in deterministic contexts. The problem is that there will always be later events that determine whether the earlier correlated events occur. Putting in different words, we might say that even when we

[8] Ibid. p.526.

[9] "Reichenbach's Common Cause Principle." Stanford Encyclopedia of Philosophy. <http://plato.stanford.edu/entries/physics-Rpcc/>, accessed on 9 February 2016.

consider a set of previous events to be correlated, it is possible that later on with the help of some later events we may come to know that the earlier events, taken to be associated, were not actually so. The PCC thus fails because it claims that typically there are no later events conditional upon which earlier correlated simultaneous events are uncorrelated.

However, this does not imply the violation of the Causal Markov Condition. In order to be able to infer causal relations from statistical ones, it is presumed that whenever unconditionally correlated quantities E_i and E_j are independent conditional upon some quantity, E_k, then E_k is a cause of either E_i or E_j. To be more precise, they assume the "Faithfulness condition[10]," according to which there are no probabilistic independencies in nature other than those which are entailed by the Causal Markov Condition. As we have seen above, since the values of such quantities X' at later times t' are not direct causes of X at t, the Faithfulness condition is violated, and the problem here is that we are unable to infer causal relations from probabilistic relations, and much of the practical value of the Causal Markov Condition is lost. Most notably, the law of *conditional independence*[11] is not violated by this type of cases discussed above.

I spelled out a few of the shortcomings of the Common Cause Principle above, and the comparison between the PCC and the CMC are only a tiny part of the vast literature available on the common cause principle. For instance, Gabor Hofer-Szabo, Miklos Redei, and Laszlo E. Szabo (2013) argue that the counterexamples provided to show the failure of the Common Cause Principle are too hasty, and are in fact ambiguous and vague. Redei (2014) further argues that evaluating the status of the PCC is, in fact, a very subtle issue. Moreover, at present, there is no meticulous evidence to prove that the PCC does not hold.

6. Conclusion

The Principle of Common Cause is a very significant tool in causal explanations. The advantage of the PCC is that correlated effects can be screened off by a common cause, because a common cause is a genuine and fundamental cause from which various effects arise. Though there are epistemological advantages of common cause explanations of correlated events, the PCC is not impeccable. I endorse the view that the PCC does qualify as a principle that can be used as a tool in explaining improbable coincidences. However, the attempt of this paper has been to show that there are scenarios in which the application of the

PCC can lead us astray, or where it might fall short. The first case is when there is an unknown common cause which would produce an effect, and that effect might appear as a common cause for other correlated events. The second case is when there are two events which are not distinct due to a spatiotemporal overlap, yet they seem to be distinct. Since event A and event B are not spatiotemporally distinct, there can be no cause that screens them off. Another occasion where the PCC might fail is when there is more than one common cause of correlated joint effects. In such cases, even if one common cause screens the effects off, the effects will still be correlated because of another common cause. Hence, though the PCC seems to be relevant in providing explanations of correlated events of various domains, I would suggest that the PCC needs appropriate restrictions, at least in the cases which have been discussed above. However, I have already granted the view that a proper evaluation of the common cause principle is a very subtle matter. Hofer-Szabo, Miklos Redei, and Laszlo E. Szabo (2013) also argue that the standard counterexamples given against the PCC are too hasty, ambiguous, and vague. Similarly, the scenarios I have pointed out above, in which the PCC falls short, may also be prone to Hofer et al's objection.

Acknowledgments: I thank the anonymous reviewer for the useful comments on the earlier version of this paper, especially for suggesting the literature on the Causal Markov Condition and its uses in cases like those described in Scenario I and scenario II.

[10] The Faithfulness condition says that all of the conditional and unconditional probabilistic independencies that exist among the variables in V are required by the MC.

[11] In probability theory, two events *A* and *B* are conditionally independent given *Y* if and only if, given knowledge that *Y* occurs, knowledge of whether *A* occurs provides no information on the likelihood of *B* occurring, and knowledge of whether *B* occurs provides no information on the likelihood of *A* occurring.

References

Arntzenius F (1990) Physics and Common Causes. *Synthese* 82: 77–96.

Arntzenius F (1993) The common cause principle. *PSA* 2: 227–237.

Arntzenius F (1997) Transition chances and Causation. *Pacific Philosophical Quarterly* 78(2): 149–168.

Hofer-Szabó G, Rédei M, László E (2013) *The Principle of Common Cause*. Cambridge: Cambridge University Press.

Hausman D M, Woodward J (1999) Independence, Invariance, and the Causal Markov Condition. *The British Journal for the Philosophy of Science*, 50: 521–583.

Hausman D (1998) Causal Asymmetries. Cambridge: Cambridge University Press.

Hitchcock C (2012) Probabilistic Causation. *Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/causation-probabilistic/index.html> accessed on 12th August 2016.

Humphreys P (1982) Aleatory Explanations Expanded. *Proceedings of the Biennial Meeting of the Philosophy of Science Association* 2: 208–223.

Humphrey, P (1989) Scientific Explanation: The Causes, some of the Causes, and Nothing but the Causes. In *Scientific Explanation*, eds Salmon W., Kitcher P., 283–304. Minneapolis: University of Minnesota.

Kitcher P (1989) Explanatory Unification and the Causal Structure of the World. In *Scientific Explanation*, 461–470. Minneapolis: University of Minnesota.

Lewis D (1986) Postscripts to ‘Causation.

Pearl J (2000) *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Redei M (2014) Assessing the Status of the Common Cause Principle, In: *The Philosophy of Science in a European Perspective*, 433–442. Springer.

Rehder B (2003) Categorization as causal reasoning. *Cognitive Science* 27: 709–748.

Reichenbach H (1956) *The Direction of Time*. Berkeley and Los Angeles: University of California Press.

Russell B (1961) *Human Knowledge: Its Scope and Value*. London: George Allen and Unwin Ltd.

Salmon W C (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Sober E (1984) Common cause explanation. *Philosophy of Science Association*, 51

Sober E (1988) The Principle of the Common Cause. In *Probability and Causality*, ed. J. Fetzer, 211–229. Dordrecht: Reidel.

Spirtes P, Glymour C, Scheines R (2000) *Causation, Prediction, and Search* (2nd ed.). The MIT Press.

Spohn W (1991) On Reichenbach's Principle of the Common Cause. In: *Logic, Language, and the Structure of Scientific Theorie*. University of Konstanz, Germany

Suppes P (1970) *A Probabilistic Theory of Causality*. Amsterdam: North-Holland

Van Fraassen B. (1982) The Charybdis of Realism: Epistemological Implications of Bell's Inequality. *Synthese* 52: 25–38.

Woodward J (2003) *Making Things Happen – A Theory of Casual Explanation*. Oxford: Oxford University Press.

Wronski L (2010) *The Common Cause Principle Explanation via screening off*, Ph.D. Thesis, Jagiellonian University, Krakow, Poland

Zhang J (2013) A Comparison of Three Occam's Razors for Markovian Causal Models. *The British Journal for the Philosophy of Science*, 64: 423–448.

Interview

HOW DOES PHILOSOPHY OF SCIENCE MAKE A DIFFERENCE IN THE WORLD WE LIVE IN? A CONVERSATION WITH STEPHAN HARTMANN, STATHIS PSILLOS, AND ROMAN FRIGG

In the Autumn of 2015, a small group of members of the European Philosophy of Science Association (EPSA) decided to found what is known today as the East European Network for Philosophy of Science (EENPS).[1] The Inaugural conference of EENPS was held in Sofia on June 24–26, 2016, and was hosted by New Bulgarian University. The conference was a great success: about 70 scholars from 22 different countries took part, and three of the founders of the EPSA accepted invitations to give keynote talks. The keynote speakers Stephan Hartmann,[2] Stathis Psillos,[3] and Roman Frigg[4] also agreed to answer some awkward questions about the mission of philosophy of science in the present world, the role of networking among philosophers for the completion of their mission, and the value which professional organizations such as EPSA and EENPS add to the lives of those who are committed to the mission of philosophy of science. The questions were asked by Lilia Gurova, who chaired the Local Organizing Committee of the Inaugural conference of EENPS;[5] the answers of Hartmann, Psillos, and Frigg follow.

Lilia Gurova: We, the founders of the East European Network for Philosophy of Science, greatly appreciate your taking part in our inaugural conference as keynote speakers. Your participation in the conference is recognized as support for our network and its aims and mission, an important part of which is to promote research and research collaboration in philosophy of science in East, South-East, and Central Europe. Many people in this region still doubt whether there are any good reasons for pursuing or supporting research in philosophy of science. I hope your answers to the following questions will help them to take an informed stance.

Q1: Philosophy of science, broadly construed, aims at a better understanding of science—of its foundations and methods, as well as of its implications for society and culture. Although many would agree that these are questions which are interesting in themselves, the majority, including some people in academia, seem to lack a clear idea of why these questions should be taken seriously. What would you tell these people?

Stathis Psillos: This kind of question is expected, especially nowadays when academic disciplines and academics themselves are “measured” by their cash-value, their short-term impact, or their “usefulness” to society at large.

But imagine asking this question to Einstein, or to Jean Perrin, or to Helmholtz, or to Henri Poincare—and before them to Newton and to Leibniz (to name but a few). If they bothered to make a case for philosophy of science, as opposed to showing by

[1] More details about the East European Network for Philosophy of Science (about its history, mission and current activities) could be found on its website: https://sites.google.com/site/eastnetworkphilsci/home

[2] Stephan Hartmann is Professor of philosophy of science at Ludwig Maximilian University (LMU) of Munich. He is Co-Director of the Munich Center for Mathematical Philosophy (MCMP), President of the European Philosophy of Science Association, and President of the European Society for Analytic Philosophy.

[3] Stathis Psillos is Professor in the Department of Philosophy and History of Science at the University of Athens.

[4] Roman Frigg is Professor in the Department of Philosophy, Logic and Scientific Method at the London School of Economics and Political Science (LSE). He is Director of the Centre for Philosophy of Natural and Social Science and Co-Director of the Centre for the Analysis of Time Series at LSE.

[5] Lilia Gurova is Associate Professor of philosophy of science in the Department of Cognitive Science and Psychology at New Bulgarian University.

their own examples its importance to, and fertility for, their own work, they would certainly start by saying that philosophy needs science as much as science needs philosophy.

This relationship of mutual dependence is constitutive of both cognitive areas and has always been so—from ancient Greek philosophy and science to fairly recent times. Science grows through conceptual changes, and philosophy—whether practiced by philosophers or scientists (or both)—is the discipline that delineates the conceptual landscape of new theories. Philosophy grows through the re-drawing of the scientific image of the world, and science is the terrain in which the scientific image of the world is developed. Philosophy of science synthesizes the scientific image of the world; at the same time, it explores its limits and its epistemological presuppositions.

Take the perennial question that I have spent some time on: can and should we trust the current scientific image of the world? Is science in the truth-business, or should it be viewed as a collection of fairy tales, soon to be replaced by others? Is science in the business of discovering what the world is like, or is it in the business of constructing reality? One might treat these questions as esoteric ones which only excite the interest of professional philosophers of science. I beg to differ. Addressing these kinds of questions is imperative for society as a whole, especially in light of the fact that science has been the driving force of the cognitive achievements of humanity as a whole. It is also equally imperative for educators and policy-makers, since what kind of curricula are designed and what kind of research projects are fostered and encouraged might well depend on how seriously we take science as a cognitive endeavor, and why.

I could add more examples in which the questions that occupy the minds of philosophers of science are not esoteric and idiosyncratic ones: for instance, is science value-free, and if not, how can there be objectivity in it? (I am sure that science is not value-free, and yet it is by and large an objective enterprise). Or another example: how should consensus be established among scientific experts, and how should this consensus be based on evidence, especially in light of uncertainty? Answering questions such as these is important in treating pressing issues like anthropogenic climate change. More could be said. Suffice it to say that the history of science teaches us that scientific progress is constitutively linked with philosophical reflection about science as a whole and the various individual sciences and theories.

Roman Frigg: There is both an intellectual and a practical motivation for engaging in philosophy of science. Over the last 300 years, science has changed our lives like no other human activity. Little, if anything, in the way we live would be the way it is without science. But what is science? What are its methods, what are its goals, and what makes it successful? These are pressing questions for everybody who appreciates the achievements of science. But science should also be looked at critically. Like every other activity, it can be done well and it can be done badly, and it is not always clear which is which. It is therefore important that we reflect on how science should be practiced, and on what counts as good and as bad science. This is not merely an academic exercise—in fact, it has direct influence on our lives. How should drugs be tested before they can be prescribed to people suffering from a disease, and how is their effectiveness assessed? Do climate models instruct us truthfully about the nature of climate change? And is the evidence in support of policy interventions sufficient to warrant large scale government programmes? All these are questions in the philosophy of science, and the answers matter to our lives.

Stephan Hartmann: Science is a fundamental part of our world and of our culture, and it is important that philosophers of science study it. It is important to learn about the implications of science for the "big questions" concerning our existence, and it is important to know how science works. What are the goals of science, and can science really achieve these goals? Is truth a goal of science, or is it only about its practical applications? Philosophers of science also study the methodology of science and critically evaluate the claim that science is a paradigmatically rational endeavor. But is science really rational? And how does the scientific approach differ from other approaches? What is the status of scientific knowledge, and what, if anything, privileges the scientific approach? These are all so-called meta-questions, i.e., philosophical questions about science. They are important, and are at the heart of our discipline for good reasons. But philosophers of science also increasingly collaborate with scientists, and these interactions have been enormously successful, as demonstrated by important work in the foundations of physics, biology, and cognitive and social science. These achievements would not have been possible without the crucial contributions of philosophers of science.

Q2: It seems to be a common belief that networking among philosophers facilitates the

spread of "news and views," but has nothing to do with the inception of deep and original ideas. Do you share this belief? Why or why not?

Roman Frigg: Of course not. Professional philosophy has nothing to do with sitting around in a cafe, smoking a filterless Gauloises, and celebrating an existentialist lifestyle. Philosophical conferences are no different from conferences in physics, biology, or geology: people work hard on papers and ideas, and then subject them to the criticism of their colleagues. This is a rigorous and methodical intellectual activity, and it makes crucial contributions to the growth of philosophical understanding.

Stephan Hartmann: In my view, there is nothing wrong about learning the "news and views" about our field and the people in it. It is important to get to know new areas of research, interesting problems and questions, funding and job opportunities, conferences and workshops, and new initiatives. The reasons for this are all too obvious. If one is not part of a network, one often misses this information (even in the age of the internet). But of course, networking has many more functions. One gets feedback on one's own work, one gets ideas after a talk from someone else and starts a collaboration, one gets new invitations, makes new friends, and in general one feels to be part of a community of like-minded people. I can say that I have profited greatly from my network in all these respects, and I do my best to help my PhD students and postdocs to build their networks.

Stathis Psillos: "Networking" is almost as old as philosophy itself. It just has different forms. In the seventeenth century, for instance, it primarily took the form of correspondence among various philosophers and laypersons, and sometimes the form of dinners in Parisian salons. In the twentieth century it took the form of conferences, congresses, and workshops. In the interwar period, for instance, a number of congresses of scientific philosophy took place in several European and American cities, and this was instrumental for the consolidation of European philosophy of science and the spread of the ideas of logical empiricism on both sides of the Atlantic. Part of the rationale for the post-WWII World Congress of Philosophy was to open up the philosophical world to philosophers from the then socialist countries, and to facilitate the spread of philosophy in them.

I could offer more examples, but the point is that networking is part of the academic world and *sine qua non* for the formation of academic communities which overcome national, cultural, and doctrinal boundaries. So the dilemma "news and views" vs "deep and original ideas" is false precisely because academic communities need both. Alfred Tarski presented his ground-breaking views about truth in a conference of the Vienna Circle in Paris in 1935—it was both news and views and deep and original ideas! And there are many cases like this. Deep and original ideas need time to mature, and exposure in conferences and similar venues is certainly an excellent way for them to grow and get spread around. Conversely, without being exposed to what's now happening in the community—the intellectual trends, the various lines of investigation, the hot topics—it's hard for someone to acquire and retain the level of creativity that is required for the production of original ideas. In any event, a lot depends on the level of "networking"—smaller and more focused conference and workshops tend to be more productive and profitable, whereas bigger and general conferences tend to promote a culture of communication and getting to know what are the current trends in the profession. In many ways, it's hard to compete for grants and other sources of funding nowadays without forming community networks and without fostering the collaboration of researchers in various countries and even continents.

Q3: You are among the founders of the European Philosophy of Science Association (EPSA), which was launched in 2007. What could you say now, almost 10 years later: Did EPSA change the philosophical landscape in Europe? What do you think is the biggest achievement of EPSA? And what has it failed to achieve (so far)?

Stephan Hartmann: One of the main reasons we founded EPSA was that we felt that we needed a venue where all European philosophers of science could meet. Something like this has existed in the US for a long time, but it was missing in Europe. Here we were very successful. EPSA organizes a large conference every other year, and many people from all sorts of countries (even from the US and Australia) attend.

Another reason to found EPSA was to have a truly European Journal for Philosophy of Science which also demonstrated to the wider philosophical world what philosophers of science are doing and why it is important and relevant. We were also successful here. The journal (*EJPS*) is well established, gets excellent submissions, and we are sure that it will not take too long until EJPS is on the same level as *Philosophy of Science* and the *British Journal for the Philosophy of Science*.

Another reason for establishing EPSA was to raise the level of work in philosophy of science in Europe. Here it is harder to tell what the relevance of EPSA actually has been, but it is clear that our field developed enormously in Europe (especially in continental Europe; the UK has always been strong) over the last ten years. I think it is fair to say that Europe is now a major player, and that there is no reason for an "inferiority complex" or "minority complex" in comparing ourselves to the US.

One thing that did not work so well is the number of members. People become EPSA members when they attend a conference, but they often do not renew their membership unless they attend the next conference. We were hoping that being a member of EPSA would be a very natural thing for PhD students and researchers in the philosophy of science. This is how it is in others fields, but we are not there yet, and I understand that other societies (even some well-established ones) have a similar problem.

Stathis Psillos: A European Philosophy of Science Association was a big dream for a number of us. There were various options available, but a few of us got together in London in 2006 and took the initiative of making the dream a reality. The prevailing thought, I can confidently say, was that philosophy of science in Europe had grown so much in quantity and quality, and that the time was ripe for showing all this to the rest of the world. But it was also an attempt to create opportunities, especially for younger colleagues, or for colleagues who come from communities with less international visibility—opportunities to present their work; to enhance their integration to the world community; to open up space for collaborations. The time was ripe indeed—in no time at all, EPSA grew to be a major international force in philosophy of science. The biennial conference became a "must go" event; the European Journal of Philosophy of Science became a "must publish" venue. The biggest achievement of EPSA is the hundreds of papers by younger researchers from all corners of Europe and around the world that have been presented in the conferences and have been published in the edited volumes that followed them. I don't think there are any significant failures. But I would wish to foster the development of a distinctive European way to do philosophy of science—in a sense analogous to the distinctively European way of doing philosophy of science in the interwar period, a way which combines formal rigor with deep philosophical sensitivity.

Roman Frigg: EPSA has brought philosophers of science in Europe together on European soil. Before EPSA, I would see my European colleagues mainly at the PSA conferences in North America. Having our own conference in Europe really brought us closer together, engendered more collaborations, and created a community that fosters good academic work. This is very important. We also started a new journal, which is coming along well and publishes good papers.

Q4: What would you like to see as a main achievement of EENPS after 10 years?

Stathis Psillos: My hope and expectation is for a renaissance of philosophy of science, and for philosophy in general, in the countries and areas represented by the East European Network for Philosophy of Science. Philosophy of science has already had a vocal presence in many of these countries, and I am very optimistic that—with the leadership of the Network—it will keep flourishing. The key is integrating philosophy of science into the students' education, both in the sciences and in philosophy in general. The Network can certainly be the driving force for bringing the various national communities in close contact, and at the same time for transgressing national boundaries and showing that philosophy is one, and that good philosophy is truly communal and truly international.

Roman Frigg: I hope that EENPS will create a support network for philosophers of science in Eastern Europe which helps philosophers from that part of Europe to connect with their other European and international colleagues. In particular, young scholars need a community in which they can work, and which provides a platform for the exchange of thoughts. Good work can only be done in a context that provides continuous intellectual stimulation and feedback, and I hope that EENPS will achieve that. The effect of it would be that we would see more Eastern European philosophers at international conferences, and that more intellectual contact with other parts of the world would be established.

Stephan Hartmann: The EENPS is an excellent and important initiative, and I am sure that it will help in (i) connecting different scholars from different (and perhaps even the same) country, (ii) raising the level of research in Eastern Europe, (iii) spurring new research initiatives for the rest of the world to pick up, and (iv) connecting Eastern and Western Europe more closely in terms of research and exchanges of ideas and scholars. Finally, (v) I hope that the EENPS will succeed in attracting more young people to our field, and that philosophy of science (and analytic philosophy in general) become more attractive for the next generation of philosophers. I wish the EENPS all the best in this!

Book reviews

A POSSIBLE WAY OF RELAUNCHING PHILOSOPHICAL CREATION IN CULTURE AND PUBLIC LIFE

[A Review of Andrei Marga***, Explorări în present (Exploring in the Present),*** **Cluj-Napoca, Eikon Publishing Housse, 2014, pp.]**

Marţian Iovan

Department of Humanities
The "Vasile Goldiş" Western University of Arad, Romania
e-mail: iovanm@uvvg.ro

Through analyzing the book *Explorări în prezent* (*Exploring in the Present*)—published by professor Andrei Marga, Cluj-Napoca, Eikon Publishing House, 2014—we could offer an alternative for raising general interest in philosophy (and contemporary philosophy in particular) among educated audiences and the intellectual elite. A great deal has been written on the philosophical preoccupations of some of the most important institutions in society: administration, justice, policy, scientific research, religion, communication, etc. We can easily see that theoretical, abstract, specialized writing in these areas has oversaturated libraries and bookshops. Hundreds of books titles in these fields sit on bookshelves waiting to be read. In comparison, applied philosophy focuses on analyzing the essence and the meaning of the changes in various segments of society (for instance, economics, politics, justice, religion, morality, science, history, communication, globalisation, identity, etc.) in order to renew interest in these more fields, both among policy makers and the general public.

Professor Marga's work belongs to this last category of philosophical writings. Actually, he also excels in the philosophical, theoretical, systematic approaches.[1] His other books testify to this fact: *Rationality, Communication, Substantiation* (1991); *The Pragmatic Reconstruction of the Philosophy* (1998); *Relativism and Its Consequences* (1998); *Introduction to Contemporary Philosophy* (2002); *The Absolute Today. Joseph Ratzinger's Theology and Philosophy* (2010); *Introduction to the Philosophical Substantiation and Methodology* (1994); *Substantiation* (2011); *The Philosophy of the European Unification* (2005); *The Changing of the World. Globalisation, Culture, Geopolitics* (2013), and others. For a change, the volume called *Explorări în actualitate* (*Exploring in the Actuality*) is part of an applied, militant philosophy which capitalizes on the author's philosophical

[1] For several decades, Andrei Marga has been a professor of logic and contemporary philosophy at the Babeş-Bolyai University of Cluj-Napoca. He was Dean of the Faculty of History and Philosophy, rector of the university, and also held the position of minister of national education, foreign affairs minister in the Romanian government, and president of the Romanian Institute of Culture. He has been a leader or member of various national and international academic organizations. His achievements in teaching, scientific research, and academic administration have been recognized through numerous international and national awards, among them the "Herder," "Bărnuţiu," and "Brăteanu" Awards.

vision in analyzing and interpreting some contemporary philosophical issues. In this context, Marga expresses the view that "the person dealing with the philosophy must enter into journalism and to give a hand to the reconstruction that followed the historical turn of 1989" (p.5).

In his published volumes of papers and essays—starting with the 1992 *Explorări în actualitate* (*Exploring in the Actuality*), and followed by such others as *The Hope of Reason* (2007) and *The Diagnosis. Articles and Essays* (2008)—Marga argues that the practical function of the philosophy in knowledge-based societies is the conceptual coordination of their evolution, allying himself with philosophers such as Hegel, Nietzsche, Husserl, and Heidegger. First, philosophical exploration is not directed toward the creation of categories, but toward investigating the happenings of the living world of humanity (p.6). However, the exploration of reality requires a critically guided philosophy. "When shortcomings are overwhelming, one can try to legitimate the critical view. According to this view, one might say that the first mature and responsible attitude toward existence relies on things as a whole. It is lucid acknowledgement of what is considered to be acceptable when thinking about the things we miss" (p.7). The thinker's adherence to universal values, the rationalist vision of the human condition, on the world and on society represents the spiritual ground for making the distinction between good and bad, functional and dysfunctional, legitimate and illegitimate, in order to identify the essential issues which the analyzed field is facing, and to establish an optimal way to solve them.

The book unites papers and essays on a wide range of themes, such as politics, law, philosophy, theology, and the function of the university. The selected issues are critically analyzed, and followed by proposed solutions to problems, and are altogether essential for social progress. They are not selected randomly, but based on author's philosophical vision and priorities. For instance, in the chapters "Current Politics," "Current Society," and "Current World," the author draws attention to various dysfunctions and obstacles; the preservation of some noxious mentalities and practices in institutional activity; the organized confusion of values; decreases in initiative, innovation, and creation; massive corruption amplifying various inequities and injustices; the failure to practice real democracy; and errors in selection of focus among policy-makers, so that the historical transition towards a society that resonates with European values, with an understanding of basic human rights, and with the rule of law seems endless.

The critical rationalism in these analyses is coordinated by a unified philosophical vision; the conception of the various problems, of the factors which generate conflict, and of bad praxis in decision-making are accompanied by solutions that are part of a larger long-term vision. Nonetheless, enacting the authors proposed concrete solutions will require specialists, experience, and a feeling of ownership and responsibility for the problems. Placing and keeping ourselves in a state of confusion over values—supplemented by a lack of ability to lead, to innovate, to create, and to motivate—constitutes a major obstacle in the balanced development of society. That is why "we should start from the characteristics of those who are making the decisions" (p.40), the author claims.

Other subjects addressed by the author include types of societies, the "Asian miracle," the crises of late modernity, globalisation, the "widening of the gaps," unemployment and poverty, power and government, foreign policy, etc. These issues require thorough diagnoses, identification of their underlying causes, and reforms in their systems of the decision-making—particularly at human resources level, by promoting qualified, able, and righteous people. The author's analyses and views rely on the concepts and ideas of a vast specialist literature, ideas from the top minds in sociology, economics, politics, anthropology, theology, etc. Numerous references are made to works of such writers as: Max Weber, John Willey, Martin Jacques, Jean-Marie Bouissou, Mathieu Duchatel, George Friedman, Michel Foucault, Gilles Chance, Joseph E. Stiglitz, James M. Davies, Erik Thorbeeke, etc. Such references mark the author's scope, his exceptional capacity to highlight what is essential, and his ability to shape the conversation around societal problems and surprise us with novel solutions; at the same time, they display the author's style, which is marked by a concern for clarity and concision, and the use of a specialized language of multidisciplinary and philosophical origin. Thus, the reader might understand the message from each author's study in the context of a synoptic presentation of the philosophical ideas circulated by the contemporary culture.

The author claims in the chapter "Current Law" that it is a mistake to consider public debate about the legal system to be disallowed as interference in judicial affairs. To the contrary, discussions about justice and the proposal of improvements to the justice system are a citizen's right, and should be considered quite ordinary to the extent that justice is considered to be at the pinnacle of universal and social values (p. 105). For instance, the American Constitution is built on guiding values-concepts: the right to life, to freedom, and to the pursuit of happiness, followed by justice. That is why achieving justice in society must not be considered a topic of discussion best limited to small minority.

The author rightfully claims that the value of the justice can be found in the laws and their implementation. However, "no legislation in the world can regulate all possible litigation." Putting aside the difficulties inherent in providing sufficient context for any given law, there are at least two other major problems interfering with the realization of justice: the good (or bad) faith of the people involved, and the literal versus applied meaning of laws (p. 105). In other words, law must be considered together with their intended meaning and applied in good faith. This is why limited positivism is mistaken when it isolates the application of laws from their immanent values. Some of the decision-makers in the judicial system and in Romanian State institutions are similarly mistaken. Providing terms that protect the wealth of both individuals and society is how the judicial system promotes justice. Such action is in direct correlation with the requirements set by democracy and the rule of law. The civic culture must stay open to debate on the political and juridical values of democracy.

The author explores the justifiably high distrust of the country's population (over 50%) when it comes to matters of justice, and the obstacles interfering in the achievement of true justice. The author emphasizes the venomous results of the political control on the judiciary. This include deficiencies in the wording of laws, so-called "wooden language" (broken language) that is used for drafting most of the legal documents in our country. This language consists of a mixture of vaguely legal terms with a superficial understanding of sociological culture, a dilettante psychology, and a narrow vision of the place and role of justice in society. It also leads to an abundance of contextual laws that are especially open to abuses, and which undergo frequent amendments. The author calls this "legislative pollution," or excess of regulation, which undermines the implementation of laws. Other problems include the excessive personalization of decisions with meagre personalization of responsibilities; the discriminatory character of appointments, both at a system level and otherwise; the massiveness of corruption among magistrates, policemen, and civil servants; the formal character of the code for professional ethics, etc.

The analysis of "legal" versus "objective" truth is interesting, and so is the distinction between justice as decided by the magistrates and their path to achieving it. The author's arguments are strengthened by ideas and principles from the works of important philosophers, sociologists, laws theorists, anthropologists, and ethicists as: Nikolas Luhmann, John Rawls, Hans Kelsen, Giorgio Del Vecchio, Ronald Dworkin, Amedeo M. Conte, James Madison, Douglas Greenberg, Jurgen von Kempski, Pascal Boniface, etc. The author succeeds in framing his opinions on law, on legal values, and on the public-private distinction within the vast context of the contemporary debates at the "top" of each field, while at the same time capitalizing on a multi-disciplinary and inter-disciplinary approach.

As a professor of contemporary philosophy, the author explores a few specific issues of philosophical culture: the unravelling of the "myth of giving" of Wilfrid Sellars, the history of philosophy as philosophy of Leo Strauss, the "identity" of Kierkegaard, the limits of the liberal society, the reestablishment of contractualism of John Rawls, the philosophy of self-awareness of Dieter Henrich, the inferentialist pragmatism of Robert B. Brandom. Obviously, the reader's comprehension of the contents and meaning of these studies can be largely facilitated by examining these philosophic works themselves, and by undertaking professor Marga's course on contemporary philosophy.

For readers familiar with Marga's work, it should be no surprise that the volume concludes with an analysis of some issues regarding universality resulting from his accumulated managerial experience as rector of the most important university in Transylvania and as the Romanian minister of national education. After the turn of 1989, the author's proclivity toward deep analyses of current theology—particularly Catholic theology—is surprising.

The chapter "Current Theology" analyzes faith, historical research on Jesus, Biblical archaeology, the historicity of Christianity, demythologisation, Jerusalem in art, the papacy of Francis, the impact of religion, and the question "Do we know Jesus?" I am persuaded that Marga oriented himself toward theological investigation due to the historical connections between philosophy and theology, which differed from one historical stage to another, ranging from neutral cohabitation, to hostility, to substitution or subordination of one to the other, to harmonization with intent to facilitate mutual development. In addition, the finding that based on the current increase in the number of believers, the return of the religion to the foreground of the contemporary culture and of mob mentality led to the increase of the priestly influence and authority. In this context arose the debate on the new relation between religion, science, philosophy, and politics. Some pontiff's actions and messages/encyclicals have far-reaching impact on the orientations of society (p. 210).

In his analyses on the role of theology and religion in contemporary society, the author capitalizes on a vast literature in the field of religious philosophy and sociology, the history of religion, "holy writings"(the Bible, the Gospel, the epistles of the apostles, the works of the Christian Church fathers), documents of the Papal institution, and the works of some great contemporary theologians. The author identifies the existence of more judicial positions in contemporary culture on the relation between philosophy and religion: Kant's successors claim that the theology must be submitted to the court of reason of philosophical critics; other thinkers, mostly theologians, claim that theology does not need philosophy; scientific atheists, represented by laic scientists and philosophers, claim that religion is meaningless trouble, and is in opposition to scientific truth, that religion should be entirely removed from people's lives and from the public scope.

The author claims that the above opinions were refuted by the evolution of philosophy and theology in recent decades (p. 263). Great contemporary philosophers such as Jurgen Habermas, John Rawls, and many others, propose the complementarity of philosophical approaches with theological ones. The same can be said of famous scientists and top theologians (including Pope Benedict XVI) who have stated that theology needs philosophy and science; the relation between them is no longer the classical one of subordination (*philosophia ancilla theologiae).* Sociologists and other researchers claim that the number of adherents to religious communities is increasing due to globalization. Globalization is accompanied by an increase in religiosity. But faith and reason may yet find each other in a relation of complementarity. The dialogue between science and religion, and between religion and philosophy, is more important than any other relation.

In the end I might added few remarks on the book's style and construction. Readers who lack philosophical training may wish that the author had spelled out his conclusions at the end of each chapter more clearly, and that he had provided an epilogue and an index of names. Further, the author is multi-lingual, and there are quotations from six languages in the volume that are not translated into Romanian. The excessive “encyclopaedism” might discourage some readers from continuing their reading. Still, the book is a great success overall, a cultural event, fully illustrating a way in which we might broaden the scope of those interested in the philosophical writing so that it again becomes a part of the wider culture.

The value of this book is great at least partly because it emerges from the author’s vast knowledge of the most essential ideas circulated by the most important contemporary philosophers, scientists, and theologians, and does its analyses in an inter- and meta-disciplinary layout. The author’s unique experience as a minister and highly respected rector gives verisimilitude and credibility to the essays in *Explorări în present (Exploring in the Present)*.