Croatian Journal of Philosophy

Vol. XX · No. 58 · 2020

Croatian Journal of Philosophy

1333-1108 (Print)
1847-6139 (Online)

Co-published by
"Kruzak d.o.o."
Naserov trg 6, 10020 Zagreb, Croatia
fax: + 385 1 65 90 416, e-mail: kruzak@kruzak.hr
www.kruzak.hr

and

Institute of Philosophy
Ulica grada Vukovara 54/IV, 10000 Zagreb, Croatia
fax: + 385 1 61 50 338, e-mail: filozof@ifzg.hr
www.ifzg.hr

Available online at http://www.ceeol.com and www.pdcnet.org

CROATIAN
JOURNAL
OF PHILOSOPHY

Vol. XX · No. 58 · 2020

Philosophy of Science

Articles

Book Reviews

Croatian Journal of Philosophy
Vol. XX, No. 58, 2020

Legitimate Mathematical Methods

JAMES ROBERT BROWN
University of Toronto, Toronto, Canada

A thought experiment involving an omniscient being and quantum mechanics is used to justify non-deductive methods in mathematics. The twin prime conjecture is used to illustrate what can be achieved.

Keywords: Mathematics, methodology, proof, thought experiment, inductive evidence.

There is a standard view of mathematics that says proofs are the one and only source of evidence and proofs are deductive derivations from first principles. This attitude has a long tradition and there is a comforting surety about it. But occasionally there are voices in opposition, including one that should be particularly influential.

> If mathematics describes an objective world just like physics, there is no reason why inductive methods should not be applied in mathematics just the same as in physics. The fact is that in mathematics we still have the same attitude today that in former times one had toward all science, namely we try to derive everything by cogent proofs from the definitions (that is, in ontological terminology, from the essences of things). Perhaps this method, if it claims monopoly, is as wrong in mathematics as it was in physics. (Gödel 1995 [1951], vol. III: 313)

I'm going to argue for the same conclusion, but I will come at it in a very different way. Instead of trying directly to liberalize the notion of evidence in mathematics, I will assume certainty in physics, that is, I will assume that the first principles of quantum mechanics (QM) are just as certain as the Peano axioms (PA), the first principles of arithmetic. The consequence for what counts as legitimate mathematical methods will surprise.

Let's begin with a parable. God parts the clouds and says: "Verily, verily I say unto you, the principles of quantum mechanics are true." Imagine God as you will. I picture Athena, goddess of wisdom and patron of science. But be sure to include her being omniscient and truthful. This means we can now know with certainty that quantum states are represented by vectors in Hilbert space; they evolve according to the

Schrödinger equation; the Born rule will give us the right probabilities for measurement outcomes; and so on. We now have perfect confidence in the truth of the standard principles of QM, which until now were merely empirically well justified. And we also know that anything we can derive from those first principles, such as Heisenberg's uncertainty principle, is unquestionably true, since logic preserves truth.

So far, so good, but we have more questions for God to answer: Is QM complete, in the sense of implying yes or no to every QM question? If P is a consequence of QM, can we derive it in a feasible time? What is the relation of QM to other theories? Do chemistry and biology reduce to QM or not? Other questions will readily come to mind.

We ask, but God won't answer. She smiles benignly then, alas, departs. (Athena frequently helped Odysseus out of a jam, then left him to fend for himself.) Suppose this is how things now stand with us. We now know much with certainty. But we remain either ignorant or only mildly confident of much else in QM. How should we proceed?

Obviously, we should try to construct derivations for as many propositions as possible. But what about the rest? We would probably continue as before. That is, we would continue with a combination of conjectures and experimental testing. Aside from the parts of QM that are clearly certain, it would be business as usual. We would continue to argue over what this involves but details would be more or less the same. There will be experimental probing, hypothesis testing, the use of various statistical techniques, thought experiments, philosophical considerations, and so on.

We would continue to tackle many problems the way we do currently. For example, perhaps there is a derivation of the details of protein folding from the principles of QM, but no such derivation can be found by humans. Calculating the energy levels of complex objects is hopelessly difficult. U_{235} is a many-body problem that can't be exactly solved. Quantum field theory is a relativistic extension of QM, not derivable from it. What about dark energy? Is this even a QM problem? God is no help in answering these questions. We have to carry on as before.

The upshot from all of this is that some physics is certain and some is not, and we will continue to learn about the latter in the same old empirical, fallible, inductive way. Why not demand certainty everywhere in QM? The argument for not doing this, if one is needed, is simple: Pre-God we have lots of justified but fallible beliefs involving QM. Then God tells us that part of this is in fact certain knowledge. Great news. Do we abandon the remaining justified beliefs on the grounds that they are not certain? No, since their status as justified but fallible beliefs remains unchanged from what it was before God certified some of it. In that respect, nothing has changed. The fact that God certifies some of it should not turn us into sceptics about the rest.

Of course, it is still debatable precisely what good scientific method is, but that is a detail that need not trouble us here. Most of QM re-

mains fallible by anybody's lights and should be investigated empirically and inductively. We have certain knowledge of the first principles of QM and their deductive consequences. The rest of QM has the same status as it had before God intervened. Does this have consequences for our knowledge claims elsewhere?

Let's turn to mathematics, where the common attitude is that much of it is certain knowledge (and we don't need God to tell us). I'll stick to an elementary part, basic arithmetic, which, for most of us, is probably as certain as anything could be.

There is a common ideology that goes along with the general attitude about mathematics. Let's assume the Peano axioms (PA), which are a set of rules characterizing the natural numbers. PA says there is a number 0, and for each number there is a successor. Thus, 1 is the successor of 0; 2 is the successor of 1, and so on. There are axioms for addition and multiplication, and for the principle of mathematical induction. These axioms are typically taken to be certainly true, or at least as certain as anything could be. Of course, there are people who claim to doubt them, but there are also people who claim to doubt the law of non-contradiction.[1]

A theorem may be asserted, according to the common ideology, if and only if there is a proof, which is a derivation from the basic axioms. (In practice a sketch of a derivation will suffice, but it is understood that that full details could in principle be provided.) Nothing else should be believed, according to this ideology — a proof is the only evidence allowed.

All of this can be easily illustrated by a famous theorem, first proved in Euclid's *Elements*. The theorem follows from PA. *Prime numbers* are numbers that cannot be factored, that is, they cannot be divided by any numbers except 1 and themselves without remainder. They include: 2, 3, 5, 7, 11, 13, … The rest are *composite numbers*, which are the product of primes. For instance, $4 = 2\times2$, $6 = 2\times3$, $8 = 2\times2\times2$, $9 = 3\times3$, $10 = 2\times5$, $12 = 2\times2\times3$, …, $2093 = 7\times13\times23$, and so on. How many primes are there?

> Theorem: There are infinitely many prime numbers.
>
> Proof: Suppose there are only finitely many primes. Hence, there is a highest p. Let $q = (2\times3\times5\times7\times\ldots\times p) + 1$. If q is a prime, then p is not the highest prime after all. If q is composite, then q is divisible by primes. But none of 2, 3, 5, …, p can divide q, since there is always a remainder of 1. Thus, some prime r must divide q. But $r > p$. Either way, p is not the highest prime. So, the initial

[1] This is perhaps unfair to those who are fictionalists, such as Field (2016) or Leng (2010). I don't wish to debate this issue here. I assume mathematical platonism or some sort of realism from the outset and argue from there. The point of this paper is not about the ontology of mathematics, but rather its legitimate epistemology. What is the best way to acquire objective mathematical knowledge, assuming there is such a thing? (Chess knowledge, by contrast, is not objective.)

> assumption that there is a highest prime is false. Thus, there are infinitely many.

Now we have two interesting systems to think about, PA and QM. The first principles of PA and QM (post God) are both certain. Anything we can derive from either we can be sure is true. And yet we treat them differently in a fundamental way. We would be happy to go beyond the certain first principles of QM and continue to use inductive methods to enlarge what we know about the physical realm. But we have been reluctant to do the same with PA. Their epistemic situations are the same, so we should have the same epistemic outlook for each.

The parallel is obvious. In the quantum case (post God), we have two kinds of propositions: (1) QM principles and logical consequences that we can actually derive and (2) all other truths of quantum mechanics that we cannot either practically or in principle derive. In the arithmetic case, we also have two kinds of propositions: (1) PA axioms and logical consequences we can derive from those axioms and (2) all other truths of arithmetic that we cannot either practically or in principle derive.

How should we respond to this schizophrenic methodological attitude? Obviously we should follow the QM example and extend our mathematical knowledge by adding various inductive techniques to PA. This will have profound implications for mathematical practice.

The twin primes conjecture will provide a good example of a more liberal way of proceeding. *Twin primes* are pairs of prime numbers of the form (p, p+2). For instance, (3,5), (5,7), (11,13), (17,19), and so on. How many are there? This is an open problem in number theory in the sense that there is no proof that the number of twin primes is either infinite or finite. Number theorists have been attacking the problem for a long time without finding the answer.[2] It is possible, of course that the problem is unsolvable, in the sense that no proof exists either way. We know from Gödel's incompleteness theorem that such unsolvable problems exist. Euler, who is often quoted on this topic, wondered about the possibility. "Mathematicians have tried in vain to discover some order in the sequence of prime numbers but we have every reason to believe that there are some mysteries which the human mind will never penetrate." (1710, quoted in Simmons 1992: 276n3).

To proceed, let's take note of the Prime Number Theorem. I will use the standard notation π(n) for the number of primes up to n, e.g., $\pi(10) = 4$. The Prime Number Theorem says: $\pi(n) \approx n/\log n$. That is, the number of primes up to some number n is approximately equal to n divided by the natural log of n. As n gets larger, the approximation becomes more accurate. For example:

[2] The literature on number theory, especially primes, is enormous. Extensive discussions can be found in Hardy and Wright (2008), Ribenboim (1991) and Shanks (1993).

$\pi(100) = 25$	$100/\log 100 = 21.7$
$\pi(1000) = 168$	$1000/\log 1000 = 144.7$
$\pi(\text{billion}) = 5084534$	$\text{billion}/\log \text{billion} = 4825494.2$

Cramér (1936) developed the idea that primes can be considered as random. If we consider them equiprobably, then the probability that a number less than n is prime is approximately $1/\log n$. The idea can be tweaked to address obvious problems (eg, half the numbers are even so not prime, aside from 2).

Think of the gap between primes. For instance, the gap between 5 and the next prime 7 is 2; the gap between 11 and 13 is also 2, while the gap between 13 and the next prime 17 is 4, and so on. The apparent randomness of the primes will be reflected in the randomness of the size of the gaps. Since there are infinitely many primes, we can expect the number of gaps of size 2 to occur infinitely often. And that means that primes of the form $(p, p+2)$ will occur infinitely often. In short, *the twin primes conjecture is true*. And it is justified by rather simple but quite compelling inductive means.

The argument is easily generalized to prime pairs of the form $(p, p+4)$, $(p, p+6)$, and so on. There are infinitely many pairs of each of these, as well, since there will be infinitely many gaps of size 4, size 6, and so on. The moral to be drawn from this example is that inductive methods can provide legitimate evidence in mathematics more generally.

I want to stress that the foregoing argument significantly differs from other arguments for inductive methods in mathematics. Besides Gödel who was quoted at the outset, lots of people (including me (Brown 2008, 2017)), have argued for such a conclusion. One of the simplest arguments for a more liberal methodology is the fact that the first principles cannot be proven (without begging the question), so it is in principle hopeless to demand that all our mathematical evidence be based on proofs. Another argument for mathematical fallibility is based on conceptual change. For instance, in the 18th century it was thought that all functions are continuous. The proof for this theorem was flawless. The concept of function, however, changed during the 19th century, so that now we take a function to be an arbitrary association between sets. This allows the radically discontinuous Dirichlet function *f(x)*, which equals 1 or 0, depending on whether x is rational or irrational.

The argument here is quite different in that it assumes that mathematics is in part certain. Specifically, the Peano axioms are taken to be as certain as anything. The argument then follows the lesson of QM resulting from the God thought experiment, namely, that inductive methods should supplement the known-to-be-certain first principles. This is why the God TE at the outset is important; it guarantees the analogy between mathematics and physics, which is the basis of the argument.

Of course, there is no God who guarantees the first principles of QM, and we cannot continue to take those principles to be certain. The thought experiment has done its job and led us to a new way of viewing legitimate mathematical methods. Now we can treat it like Wittgenstein's ladder. Toss it out and agree that even the first principles of QM and PA are fallible, as is all knowledge, but the liberalization in what counts as evidence more than makes up for the loss of certainty.

Acknowledgements

Thanks to the audience in Dubrovnik (April 2019), especially: Mary Leng, Richard Dawid, and Zvonimir Šikić.

References

Brown, J. R. 2008. *Philosophy of Mathematics: Contemporary Introduction to the World of Proofs and Pictures*. 2nd ed. New York: Routledge.

Brown, J. R. 2017. "Proofs and Guarantees." *Mathematical Intelligencer* 39: 47–50.

Cramér, H. 1936. "On the order of magnitude of the difference between consecutive prime numbers." *Acta Arithmetica* 2: 23–46.

Field, H. 2016. *Science Without Numbers*. 2nd ed. Oxford: Oxford University Press.

Gödel, K. 1995 [1951]. *Collected Papers*. Oxford: Oxford University Press.

Hardy, G. H. and Wright, E. M. 2008. *Introduction to the Theory of Numbers*. Oxford: Oxford University Press.

Leng, M. 2010. *Mathematics and Reality*. Oxford: Oxford University Press.

Ribenboim, P. 1991. *The Little Book of Big Primes*. New York: Springer.

Shanks, D. 1993. *Solved and Unsolved Problems in Number Theory*. New York: Chelsea.

Simmons, G. 1992. *Calculus Gems*. New York: McGraw-Hill.

Croatian Journal of Philosophy
Vol. XX, No. 58, 2020

The Effectiveness of Representations in Mathematics

JESSICA CARTER
University of Southern Denmark, Odense, Denmark

This article focuses on particular ways in which visual representations contribute to the development of mathematical knowledge. I give examples of diagrammatic representations that enable one to observe new properties and cases where representations contribute to classification. I propose that fruitful representations in mathematics are iconic representations that involve conventional or symbolic elements, that is, iconic metaphors. In the last part of the article, I explain what these are and how they apply in the considered examples.

Keywords: Visual representations, discovery, iconic metaphors, manipulation, Peirce.

1. *Introduction*

Many scholars have commented on the advantages for mathematics of choosing appropriate notations. Euler, for example, expressed that Leibniz's notation for the differential was superior to Newton's:

> It might be uncivil to argue with the English about the use of words and a definition, and we might easily be defeated in a judgment about the purity of Latin and the adequacy of expression, but there is no doubt that we have won the prize from the English when it is a question of notation. For example, the tenth differential, or fluxion, is very inconveniently represented with ten dots, while our notation, $\mathrm{d}^{10}\mathrm{y}$, is very easily understood. (Euler 2000: 64)

Other mathematicians comment on the potential "fruitfulness" of a good choice of notation:

> It only becomes possible at all after the mathematical notation has, as a result of genuine thought, been so developed that it does the thinking for us, so to speak. (Frege 1953: xvi)

Another concern is the choice of representations in mathematics. A recent such interest is the role of visual representations, or diagrams. The aim here is to show particular ways in which visual representations contribute to the development of mathematical knowledge. One focus will be to illustrate how these representations enable you to *see* or observe certain patterns which leads to the formulation of new hypotheses. A puzzle that I will address—but only partially solve—concerns the question of *how* and *why* certain representations contribute to the development of mathematics. One part of the answer (see Carter 2019) is that it is often fruitful to have available *iconic* representations that are possible to *manipulate*. Taking as a starting point Peirce's characterisation of an icon, I will first propose that icons used in mathematics are best understood as iconic metaphors and explain what this means. In this context, I will note that iconic representations that can be manipulated play a key role in Peirce's characterisation of mathematical reasoning. Second, I will indicate that we still lack an account of how to find a useful representation or notation in mathematics.

The use of visual representations and notations has contributed to the development of mathematics in various ways. Sometimes the choice of a particular notation enables one to *see* that there is a problem of a certain type. As an example, I could mention Descartes' convention of writing a^2 instead of $a \cdot a$, a^3 instead of $a \cdot a \cdot a$ and so on. This made him able to write, for the first time, a quadratic equation (almost) as we do today, for example as $ax^2 + bx = c$. This convention made it possible to formulate a general n-degree equation and formulate the Fundamental Theorem of Algebra. As is noted by Manders this invention also suggested to Descartes why the classical problems of duplicating a cube and trisecting an angle by ruler and compass were impossible to solve:

> First, its degree, algebraically the key feature. Descartes guesses that the degree determines by what means solutions may be constructed, e.g., because angle trisection problem gives an irreducible third-degree equation, it cannot be done by ruler and compass. But there is no direct way to predict the degree of its equation from the appearance of a geometrical figure. (Manders 1989: 558).

Descartes was able to translate, for example, the problem of duplicating a cube into the cubic equation $z^3 = 2b^3$. Given a cube with side b and volume b^3, z corresponds to the side of the cube that has two times this volume. Having found that roots of quadratic equations *could* be constructed by ruler and compass, Descartes formed a hypothesis that this could not be the case for irreducible cubic equations. He also formed what he thought was a proof of this. But it turned out not to be correct. See (Lützen 2010) for details. Descartes did not yet have the required algebraic tools, for example, field extensions and formulated a geometric proof.[1]

[1] Lützen (2010) remarks that it is not strange that Descartes formulated a geometric proof: There was a long tradition of giving geometrical proofs at the time, combined with the fact that algebra was still in its infancy and so not considered as trustworthy.

Another example concerns how a particular choice of representation of a problem contributes to *classification:* a particular representation may help one to formulate—and solve—all problems of a particular type in a systematic way. The Arabic mathematician Al-Khwarizmi (c. 780–850) formulated quadratic equations in terms of the "three types of numbers" roots (the unknown), squares and numbers.[2] One of these types of equations is 'Square and roots is equal to a number'. Perhaps these expressions and their geometrical representations, when demonstrating their solution, helped him to formulate all types of quadratic equations. In any case, one usually attributes to the Arabic mathematicians the first systematic solution of quadratic equations. Other examples of representations contributing to a classification of a type of objects can be found in (Eckes and Giardino 2018).

In the next section we shall see that these two roles of representations also occur in contemporary mathematics. That is, one finds examples of representations that enable one to *see* certain properties and cases where representations contribute to *classification.* In both examples the representations consist of diagrams.

2. *Visual representations in contemporary mathematics*

Representations in free probability theory—seeing

It is possible to find examples from contemporary mathematics where a specific form of representation has contributed to the formulation of new hypotheses. One such example is presented in Carter (2010). This example illustrates how the visual appearance of a particular representation may lead to the formulation of a new concept. The example has to do with permutations on the set $\{1,2,3,\ldots,p\}$ which appear in a certain combinatorial expression in free probability theory. By representing these permutations in a certain way, certain properties of them became visible. Similar representations further contribute to make visible that these properties have an effect on the value of the expression.

The expression and its value is $\mathbb{E} \circ Tr_n\left[B_1^* B_{\pi(1)} \cdot \ldots \cdot B_p^* B_{\pi(p)}\right] = m^{e(\hat{\pi})} \cdot n^{o(\hat{\pi})}$. The B_i's in the expression stand for $m \times n$ matrices and their entries are Gaussian random variables. After taking the trace of the multiplied matrices, it therefore makes sense to take the expectation, '$\mathbb{E}$'. The indices contain 'π' which denotes a permutation on the set $\{1,2,3,\ldots,p\}$. A permutation is a 1–1 and onto function on a set to itself. The numbers, $o(\hat{\pi})$ and $e(\hat{\pi})$, in the above formula refer to the number of odd and even numbers, respectively, of certain equivalence classes on the set $\{1,2,3,\ldots,2p\}$. The total number of equivalence classes turns out to depend on properties of the permutation. I will show the representation of permutations that revealed this property. Representing a permuta-

[2] Al-Khwarizmi formulates and solves six different problems, for example, the problem 'square and roots identical to number' and 'square and number identical to roots', see (Berggren 1986).

tion by certain diagrams gives rise to the concept of a 'non-crossing permutation'. See (Carter 2010) or (Haagerup and Thorbjørnsen 1999) for further details about the case.

Below are two examples of representing a permutation on the set $\{1, 2, 3, 4, 5, 6\}$. In the diagram on the left in figure 1, you may observe that the lines do not cross, whereas they do in the right-hand diagram. This gives rise to the notion of a non-crossing and a crossing permutation.

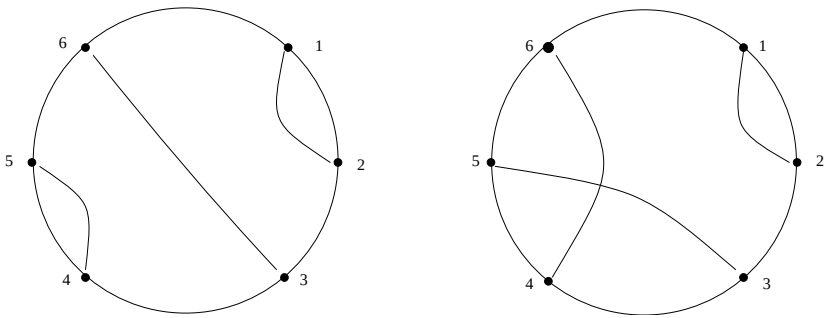

Figure 1. *The left diagram is a representation of a non-crossing permutation. In two-cycles, the permutation can be written as (12)(36)(45). The diagram on the right shows a crossing permutation. The represented permutation in this case is (12)(35)(46).*

It turns out that the above-mentioned result depends on whether lines cross or not, that is, whether the permutation is crossing or not. To see this, the mathematicians visualised, or represented, equivalence classes of an equivalence relation formed on the set $\{1,2,3,\ldots,2p\}$. (First the permutation is rewritten, taking into account that there are $2p$ matrices in the expression. The new permutation is denoted $\hat{\pi}$.) The relation is $i \sim \hat{\pi}(i) + 1$ (mod 2p). Representations of such equivalence classes can be seen below in figure 2.

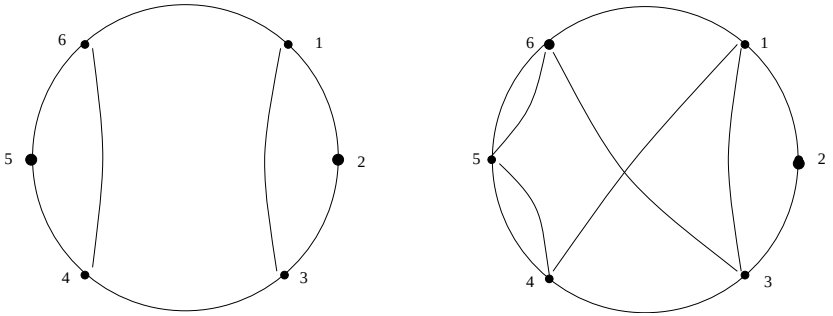

Figure 2. *Numbers that are in the same equivalence class are joined by lines. I have identified the equivalence classes of the two permutations shown in figure 1. In the left figure one sees that the number 1 is related to* $\hat{\pi}(1) + 1 = 3$. $\hat{\pi}(1)$ *is seen to be 2 in the left-hand diagram in figure 1. Similarly,* $\hat{\pi}(3) + 1 = 6 + 1 = 1$ *(mod 6), so {1,3} form one equivalence class. It can be seen that 2 is related to itself, so there is only one number in this equivalence class (marked by a filled circle). It is seen that there are 4 equivalence classes in the left-hand diagram, whereas there are only 2 in the right-hand diagram. Recall that this corresponds to the crossing permutation.*

By drawing a number of such diagrams, varying the permutation, it is possible to detect a pattern. If $p = 3$ and so $\hat{\pi}: \{1, 2, \ldots 6\} \rightarrow \{1, 2, \ldots, 6\}$ one will see that whenever the permutation is non-crossing, there are 4 equivalence classes. If the lines cross, there will be fewer. In general, the mathematicians were able to formulate the hypothesis, that the total number of equivalence classes depends on whether the permutation is crossing or non-crossing: If it is non-crossing, the number of equivalence classes is $p+1$. If it is crossing, this number will be strictly less.

In the published papers presenting this result, there are no diagrams. In order to formulate these propositions and proofs of them, the property of being a crossing permutation therefore had to be reformulated. The formal definition of a crossing permutation is as follows: A permutation $\pi: \{1, 2, \ldots, p\} \rightarrow \{1, 2, \ldots p\}$ has a crossing, if for some $a<b<c<d$ in $\{1, 2, \ldots, p\}$ it is the case that $\pi(a) = c$ and $\pi(b) = d$. If it has no crossings, it is said to be a non-crossing permutation.

One point is that there is a difference in how we *perceive* these definitions. In the diagrams the properties are *shown*. One can actually perceive the lines crossing. In the formal mathematical language, we cannot see this directly. The definitions of these properties are only *described*. (See Carter 2019 for an elaboration of this point.) Note also that this example illustrates Manders' point; that a different representation may reveal new properties or explanations. Whereas Manders discusses an algebraic representation of geometrical figures, the example presented here conversely considers a representation of a formal expression.

Representations in analysis—classification

In analysis, one field studied concerns C^*-algebras and their classification. That is, having defined C^*-algebras, one wishes to figure out the different types of such objects there are up to isomorphism. A tool to do that is to define so-called invariants. The mathematician George Elliott has formulated a program where the hope is that *K*-theory could provide such a tool: That two C^*-algebras are isomorphic if and only if their corresponding *K*-groups are pairwise isomorphic. This turned out only to be true in simple cases. The study of their *K*-groups, however, is still an important field of study. For C^*-algebras it is possible to define two such groups, denoted K_0 and K_1. It is generally quite complicated to calculate these groups from their original definitions. Recently a much easier way to calculate them has been found. The trick is first to represent the algebras in a different way, as directed graphs. From this representation, it is possible to find a different way to access these groups. I give a few details of these concepts here before coming to the main (philosophical) points: That certain diagrammatic representations are used as tools for classifying C^*-algebras and that these diagrams can be manipulated.

A directed graph is defined by a four-tuple, $E = (E^0, E^1, r, s)$. Here E^0 consists of the vertices of the graph and E^1 consists of the edges. That E is a *directed* graph means that edges have a direction, which is expressed by a *range* and a *source* function. For each edge, these functions say where it ends and starts: $r, s: E^1 \to E^0$. An example of a (finite) graph is given in figure 3. This graph has three vertices, named v_1, v_2 and v_3, and three edges, e_1, e_2 and e_3. The arrows indicate their source and range. The source of the first two is v_1, the source of e_3 is the vertex v_2. The ranges are given as follows: $r(e_1) = v_1, r(e_2) = v_2$ and $r(e_3) = v_3$.

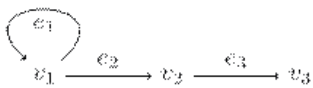

Figure 3. *A directed graph, E, with three vertices.*

A directed graph gives rise to certain generators and relations that the generators must fulfil, which then generate a C^*-algebra. The C^*-algebra generated by the graph, *E,* is denoted $C^*(E)$.For details of how such algebras are constructed, see Szymanski (2002). Read in a different way, a graph gives rise to a linear map, $\Delta: \mathbb{Z}V \to \mathbb{Z}E^0$, where *V* is the set of vertices that emit edges. It turns out that the two *K*-groups can easily be calculated from this map. First, the linear map is defined on vertices, v, that emit edges as

$$\Delta_E(v) = (\sum_{s(e)=v} r(e)) - v.$$

The two groups K_0 and K_1 can be calculated as the cokernel and kernel of this map:

$$K_0(C^*(E)) \cong coker\ (\Delta_E)$$

and

$$K_1(C^*(E)) \cong ker\ (\Delta_E).$$

In the case of quadratic equations, I suggested that the geometric representation of them contributed to the formulation of, and solution to, all types of such equations. In other words: a classification of quadratic equations. The directed graphs can be used as *tools* for classification. But they are not themselves objects of such a classification in the sense that two different graphs correspond to two different types of C^*-algebras. To a particular directed graph corresponds a linear map from which the proposed invariants, K_0 and K_1 can be obtained. Furthermore, two different graphs will give rise to different linear maps. But unfortunately, the information obtained from the K-groups is not always sufficient to tell whether the corresponding C^*-algebras are isomorphic or not. The graphs are epistemic tools in the sense that they have made calculations of the K-groups easier (Carter 2018).

Another point is that the directed graphs can be manipulated. In order to illustrate this point, we consider a result from (Szymanski 2002). It is proven that a large class of C^* −algebras can be generated by directed graphs—and so that their *K*-groups can easily be calculated. This result has been found by manipulating directed graphs. The result states that, given two specific groups, K_0 and K_1, it is possible to construct a directed graph, *E,* such that the C^* −algebra it generates has these two as its K_0 and K_1-groups, that is, $K_i(C^*(E)) \cong K_i$ for i=0 and 1. The proof—and the way this result was found—starts by considering a particular graph that gets the result partially. That is, the first graph has the right K_0-group but the other group is zero. After that a number of subgraphs are added, so one gradually gets closer to the sought for graph. One adds vertices and edges and along the way calculates how these changes alter the K-groups. Manipulating graphs, i.e., adding and removing edges and vertices, therefore contributed to the result in question.

Manipulating iconic representations

We now address the observed similarities of the two case studies. In both cases certain objects are represented by diagrams. In the first case, the objects represented are permutations and, in the second, C^*-algebras. In the first case the visual representation contributed with a new concept (that of a crossing permutation). The second example is slightly different—the representation has made progress possible because calculations of K-groups turned out to be much easier. In both

cases particular instances of concepts, that is particular examples of permutations and C^*-algebras, can be represented by diagrams. One reason that these representations contribute to new knowledge, is the fact that they can be manipulated. In this way they become tools for experimentation. By, for example, producing a number of examples of permutations and their equivalence classes one is able to detect a general pattern: that this number depends on the visual appearance of the lines in the diagram.

Another key feature of a fruitful representation is that it shares relevant "structure" with the problem, it represents. In C.S. Peirce's semiotics such representations are referred to as *icons*. An icon is the particular type of sign that is able to represent its object because it is like this object in some respect. This also entails that an iconic sign should hold the capacity to reveal more information about the object it represents, than is required to identify it as a representation of that object. Stjernfelt (2007) refers to this feature as the 'operational account' of similarity, and so of an icon. Simple examples of iconic representations consist of images and, in mathematics, of geometrical figures. These representations visually resemble what they represent. According to Peirce, icons play a key role in mathematics in general. But mathematical icons are rarely simply pictures of what they stand for. This means that the likeness must consist of something else besides visual resemblance. When Peirce characterises icons, he sometimes refers to them as having conventional (i.e. symbolic) features or that they have a purpose:

> For example, a geometrical figure drawn on paper may be an icon of a triangle or other geometrical form. If one meets a man whose language one does not know and resorts to imitative sounds and gestures, these approach the character of an icon. The reason they are not pure icons is that the purpose of them is emphasized. A pure icon is independent of any purpose. It serves as a sign solely and simply by exhibiting the quality it serves to signify. (Peirce 1998: 309)

Note that, according to Peirce, not even a drawn geometrical figure is a pure icon. I therefore propose that icons used in mathematics contain conventional, or symbolic elements—and so cannot be pure icons. They are what he refers to as *iconic metaphors* (*Collected Papers* 2.277). A related point is that, according to Peirce, a sign must be interpreted as a sign in order to function as such. To identify in which respect a sign stands for another mathematical object is therefore part of the role of the interpretant of a sign. The conventional element of an iconic sign or, in other words, the information given so that one may identify how a given sign stands for another object, I will refer to as formulating *the underlying convention or rule for interpretation*. I propose that it is a combination of (what follows from) the underlying conventions and properties of the representation that contributes to the successful use of iconic representations in mathematics.

To give a simple example of an iconic metaphor, I return to the second example mentioned in the introduction. The particular exam-

ple concerns the geometric representations of quadratic equations. One of the problems formulated by Al-Khwarizmi was 'A square and 10 of its roots is 39'. Using contemporary notation, we can also write: $x^2 + 10x = 39$. When forming a geometric representation of this problem we could formulate the following conventions: (1) both sides of the equality sign denote (the area of) geometrical figures, (2) 'x' and 10 refer to (the length of) line segments, (3) addition means that the geometrical figures are joined, (4) multiplication of two line segments gives a rectangle (or a square). These lead to a representation of the equation as shown in figure 4. This geometric figure can be manipulated to determine the line segment, x. I speculate that it is easier to obtain the solution of the equation by these manipulations than manipulating the corresponding expression or equation. It appears at least to be the way that the solution was originally found: Al Khwarizmi is said to have been inspired by Babylonian mathematicians. According to (Høyrup 2002) they solved such equations geometrically. The steps are shown in figure 5. One first cuts off half of the rectangle and moves it below the figure as shown in figure 5. In the next step, the "square is completed": one adds a square with area $5 \cdot 5 = 25$, so that the area is now 64. The side of the square is then 8 and the sought for line segment is 8–5=3.

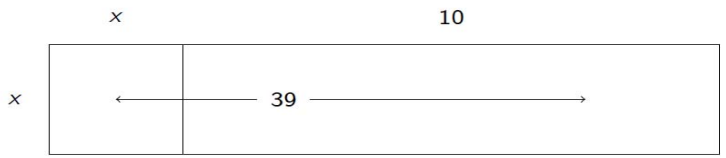

Figure 4. *A geometric representation of* $x^2 + 10x = 39$.

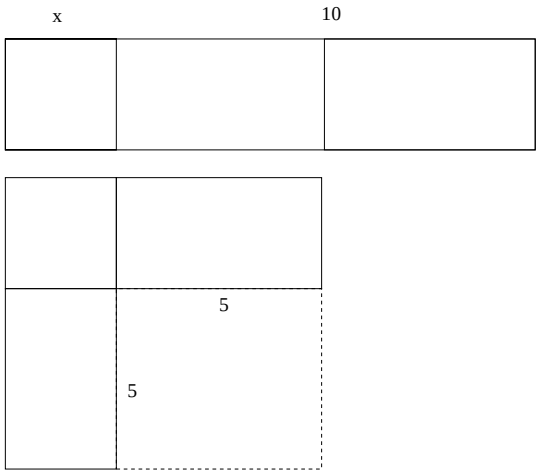

Figure 5. *Illustrating the geometric solution of* $x^2 + 10x = 39$.

Once the solution has been found geometrically, it is possible to formulate the manipulations of the figures in Figure 5 in the original language: The top figure expresses that $x^2 + 10x = 39$ is the same as $x^2 + 5x + 5x = 39$. The stippled lines in the bottom figure state that: $x^2 + 5x + 5x + 5 \cdot 5 = 39 + 5 \cdot 5$. The figure further shows that this is a square with side $5 + x$, that is, $(5 + x)^2 = 64$. Finally, one takes the square root and subtracts 5 to obtain $x = 3$.

Given this terminology, we can say that a C^*-algebra is an iconic metaphor of a directed graph. There are specific rules that define how to read a particular graph. Similarly, other definitions say how to read the graph in a different way and so obtain the linear map. This means that the linear map is also a metaphorical representation of the directed graph. Intricate mathematical arguments are needed in order to determine the relation between this map and the K-groups referred to above.

It is much easier to comprehend how diagrams represent permutations as was shown in the first case study. The employed convention is simply to place numbers on a circle and to draw a line between the numbers i and $\pi(i)$ of a given permutation, π. By using this convention, one may consider these diagrams as iconic representations of permutations. After manipulating such diagrams, the discovered property of being a crossing permutation can be reformulated in the original vocabulary of the permutation as a mapping.

The examples shown illustrate that the manipulation of iconic representations is a fruitful practice in mathematics. This brings me to the final point of this paper: that both these features play a central role in Peirce's characterisation of mathematical reasoning. In 'On the algebra of logic. A contribution to the philosophy of notation' Peirce writes the following about reasoning, mentioning the role of icons and our ability to manipulate them:

> The truth, however, appears to be that all deductive reasoning, even simple syllogism, involves an element of observation; namely, deduction consists in constructing an icon or diagram the relations of whose parts shall present a complete analogy with those of the parts of the object of reasoning, of experimenting upon this image in the imagination, and of observing the result so as to discover unnoticed and hidden relations among the parts. ... As for algebra, the very idea of the art is that it presents formulae which can be manipulated, and that by observing the effects of such manipulation we find properties not to be otherwise discerned. (Peirce in *Collected Papers* 3.363)

In this paper I have emphasised the role of visual representations, or diagrams. But it is clear from the above quote, that also other types of representations, that is, general mathematical expressions, are examples of iconic representations that can be manipulated—and so contribute to the development of mathematics.

3. *Conclusion*

I have shown various examples illustrating the effectiveness of visual representations in contemporary mathematics. In the first example a particular diagrammatic representation revealed new properties of a permutation. In the second example, a diagrammatic representation has contributed with tools that potentially make classification of C^*-algebras simpler.

I have also noted that fruitful representations in mathematics are iconic metaphors that can be manipulated. Furthermore, such representations need not be visual or diagrammatic. Finally, I should say that what has been formulated here is only a proposal of what kinds of representations are effective. The question of how they can be found remains.

References

Berggren, J. L. 1986. *Episodes in the mathematics of medieval Islam*. New York: Springer.

Carter, J. 2010. "Diagrams and Proofs in Analysis." *International Studies in the Philosophy of Science* 24: 1–14.

Carter, J. 2018. "Graph-algebras—faithful representations and mediating objects in mathematics." *Endeavour* 42: 180–188.

Carter, J. 2019. "Exploring the fruitfulness of diagrams in mathematics." *Synthese* 196: 4011–4032.

Eckes, C. and Giardino, V. 2018. "The Classificatory Function of Diagrams: Two Examples from Mathematics." *In Diagrammatic Representation and Inference—0th International Conference, Diagrams 2018*. New York: Springer: 120–136.

Euler, L. 2000. *Foundations of Differential Calculus*. Translated by John D. Blanton. New York: Springer-Verlag.

Frege, G. 1953. *The Foundations of Arithmetic*. Translated by J.L. Austin. Harper Torchbooks. New York: Harper and Brothers.

Haagerup, U. and Thorbjørnsen, S. 1999. "Random matrices and K-theory for exact C^*-algebras." *Documenta Mathematica* 4: 341–450.

Høyrup, J. 2002. *Lengths, Widths, Surfaces: An Examination of Old Babylonian Algebra and Its Kin*. New York: Springer.

Kumjian, A. et al. 1997. "Graphs, groupoids, and Cuntz-Krieger algebras." *Journal of Functional Analysis* 144: 505–541.

Lützen, J. 2010. "The algebra of geometric impossibility: Descartes and Montucla on the impossibility of the duplication of the cube and the trisection of the angle." *Centaurus* 52 (1): 4–37.

Manders, K. 1989. "Domain extensions and the philosophy of mathematics." *The Journal of Philosophy* 86 (10): 553–562.

Peirce, C. S. (1931–1960). *Collected Papers of Charles Sanders Peirce*. Vol I–IV. Edited by Charles Hartshorne and Paul Weiss. Cambridge: The Belknap Press of Harvard University Press.

Peirce, C. S. 1998. *The Essential Peirce. Selected Philosophical Writings*. Volume 2 (1893–1913). Edited by the Peirce Edition Project. Bloomington: Indiana University Press.

Stjernfelt, F. 2007. *Diagrammatology. An Investigation on the Borderlines of Phenomenology, Ontology and Semiotics.* Synthese Library (336). Dordrecht: Springer.

Szymanski, W. 2002. "The range of *K*-invariants for *C**-algebras of infinite graphs." *Indiana University Mathematics Journal* 51 (1): 239–249.

Croatian Journal of Philosophy
Vol. XX, No. 58, 2020

Mathematics and Physics within the Context of Justification: Induction vs. Universal Generalization

MARKO GRBA and MAJDA TROBOK
University of Rijeka, Rijeka, Croatia

Motivated by the analogy which holds within the context of discovery between mathematics and physics, we aim to show that there is a connection between two fields within the context of justification too. Based on the careful analysis of examples from science (especially within the domain of physics) we suggest that the logic of scientific research, which might appear as enumerative induction, is deduction, and we propose it to be universal generalization inference rule. Our main argument closely follows the analysis of the structure of physical theory proposed by theoretical physicist Eugene P. Wigner.

Keywords: Mathematics-physics analogy, context of justification, enumerative induction, universal generalization, Wigner's account of physics.

1. *Introduction—context of discovery vs. context of justification*

While it might seem unproblematic to defend the view that the analogy between mathematics and the natural sciences holds in the context of discovery, the idea to expand such an analogy to the context of justification seems to be far more problematic. We shall first introduce some preliminaries, that we shall take for granted in this paper and then present our main thesis.

We take the underlying ontology in the philosophy of mathematics to be a version of (standard) platonism but platonism in the philosophy of mathematics won't be discussed in this paper. The development of mathematical knowledge as well as the process of discovery in the natural sciences can be standardly analysed from different perspectives:

we might decide to opt for the cognitive science orientated research or a computationally orientated research, or a historically orientated research. In the context of the examples we further analyse we find the historically orientated research most appropriate.

Within the descriptive epistemic context there were offered three main epistemic routes: (1) perception—both visual and platonic; (2) experimentation and (3) positing (Trobok 2018). For each of these epistemic paths in mathematical research Trobok shows there is a counterpart in the domain of research in the natural sciences. As far as the underlying logic within the context of discovery goes, it is important to underline the difference between formal proofs in mathematics and the heuristic explanatory and exploratory procedures. Lakatos emphatically stresses the difference between formal proofs in mathematics and the heuristics of mathematical discovery (Lakatos 1976). Once such a distinction is brought to surface, the heuristics of mathematics and that in physics turn out to be analogous (Trobok 2018).

The question at this point is: How and to which extent, if at all, could the analogy (that holds between mathematics and the natural sciences in the context of discovery) be expanded to the context of justification? Namely, as Pòlya underlines:

> …many mathematical results were found by induction first and proved later. Mathematics presented with rigor is a systematic deductive science but mathematics in the making is an experimental inductive science. […] In mathematics as in the physical sciences we may use observation and induction to discover general laws. But there is a difference. In the physical sciences, there is no higher authority than observation and induction but in mathematics there is such an authority: rigorous proof. (Pòlya 1945: 117)

While in the context of discovery of both mathematics and physics we have reasons to accept Pòlya's view (Trobok 2018), the aim of this paper is to go one step further and show why Pòlya's view within the context of justification is not accurate. The aim is to show that, not just the two domains are analogous within the descriptive epistemic context, but that the analogy could be expanded to the context of justification as well. When focusing on the context of justification, we shall confine our research to the third epistemic path as above presented: *the experiment*.

2. *Context of justification specified*

In the domain of mathematics, Frege nicely explains what characterises the context of justification in his famous *Grundlagen* paragraph:

> …it is in the nature of mathematics always to prefer proof, where proof is possibile, to any confirmation by induction. […] The aim of proof is, in fact, not merely to place the truth of a proposition beyond all doubt, but also to afford us insight into the dependence of truths upon one another. After we have convinced ourselves that a boulder is immovable, by trying unsuccessfully to move it, there remains the further question, what is it that supports it so securely? (Frege 1884/1967: §2)

Let us analyse more closely the mainstream view regarding the difference between mathematics and the natural sciences within the context of justification according to which:

> The status of mathematical knowledge […] appears to differ from the status of knowledge in the natural sciences. The theories of the natural sciences appear to be less certain and more open to revision than mathematical theories. (Horsten 2017)

We shall try to show such a view being flawed by concentrating our line of argumentation on the notion of experiment. The standard view is that experiments belong to the empirical sciences, i.e. to the sphere of practical research. Experiments are practical procedures generally done by researchers in laboratories. Hence, what happens in experimental science might seem at first sight to be remote from the standard mathematical practice, mathematics being an armchair activity. And even if someone like Putnam (Putnam 1979: xi) admits that there are mathematical procedures that could be labelled as experiments (e.g. the adoption of the axiom of choice), experiments do not belong to the mathematical domain.

Standardly, experiments play several roles in science: we use them to test theories, to call for a new theory, to help us determine the structure or mathematical form of a theory, or to provide evidence for the entities involved in a theory (Franklin and Perović 2019). They hence belong to the intersection of the context of discovery and the context of justification. At first sight, someone might complain that experiments are practical procedures done in laboratories and that nothing in the mathematical domain can be analogous to such procedures, especially given the *a priori* nature of mathematical research. A closer analysis of the concept of experiment will show us, though, that the way we are accustomed to perceiving experiments does not correspond with neither the nature nor the role that experiments have played throughout the history of natural sciences (especially physics).

Galileo Galilei, the *father* of experimental physics, includes in his (Galilei 1638) the taxonomy of experiments. There are, according to Galileo, three types of experiments: real, imaginary and thought experiments. The real are those that have been performed in practice, the imaginary are those that could have been performed but haven't yet been, while the thought experiments are those that could not possibly have been performed due to the lack of technology or because impossible in principle. What is of interest to us is the fact that thought experiments are not marginal for the development of physical theories. Quite the contrary, such experiments have played a major role in the development of scientific theories in the work of Galileo, Newton, Einstein, Heisenberg *et al*. Let us mention some of the most famous thought experiments: Galileo's experiment with the result that all bodies fall at the same speed, Maxwell's demon, Einstein chasing a light beam, the twins paradox, Heisenberg's microscope, Schrödinger's cat.

Some of those experiments[1] are analogous to deductive mathematical proofs, so the analogy between the empirical physics and the *a priori* mathematics reveals itself to be of quite an importance. Let us have a closer look at the Galileo's experiment with the result that all bodies fall at the same speed (Galilei 1638).

Galileo proved, by using a thought experiment, Aristotle's theory of gravity to be flawed. According to Aristotle's theory, objects fall at the speed directly proportional to their mass. More than seventeen centuries later, Galileo writes:

> Aristotle says that "an iron ball of one hundred pounds falling from a height of one hundred cubits reaches the ground before a one-pound ball that has fallen a single cubit." I say that they arrive at the same time. (Galilei 1638/1914: [109])

The proof he offers is the following one: Galileo imagines two bodies H and L, one (H) heavier than the other (L), that are attached one to another. According to Aristotle, the compound body (H + L) falls faster than the body H, since the compound body is heavier. It means that the velocity of the united bodies is bigger than the velocity of the heavier one: $v(H + L) \geq v(H)$. On the other hand, as Galileo nicely explains:

> … when the small stone moves slowly it retards to some extent the speed of the larger, so that the combination of the two, which is a heavier body than the larger of the two stones, would move less rapidly … (Galilei 1638/1914: [109])

It follows that the velocity of the compound body should be smaller than the velocity of the H body: $v(H + L) \leq v(H)$. From the two equations it follows mathematically that the two velocities are equal: $v(H + L) = v(H)$. Galileo's result follows *deductively* from Aristotle's presumptions. Even though thought experiments clearly can serve as examples of deductive proofs in physics, such results are often treated as exceptions. The mainstream view being that:

> … the methods of investigation of mathematics differ markedly from the methods of investigation in the natural sciences. Whereas the latter acquire general knowledge using inductive methods, mathematical knowledge appears to be acquired in a different way: by deduction from basic principles. […] The status of mathematical knowledge also appears to differ from the status of knowledge in the natural sciences. The theories of the natural sciences appear to be less certain and more open to revision than mathematical theories. (Horsten 2019)

3. *Induction vs. universal generalization*

Are thought experiments marginal exceptions to the standard methods of discovering the laws of physics (science), and is the view that *in the*

[1] On the other hand, some of the thought experiments might be viewed as examples of inductive logic as advocated in (Norton 1991), but we would argue that in those examples as in the examples of real experiments, deductive rule of universal generalization is at work.

physical sciences, there is no higher authority than observation and induction (Polya 1945: 117) the right view?

Let us start with another experiment, this time from chemistry.[2] In 1828 Friedrich Wöhler was trying to synthesize ammonium cyanate from silver cyanate and ammonium chloride and obtained a white powder which he suspected was not the desired compound but could not test it as it was not obtainable in the pure enough form. He tried a different pair of chemicals, lead cyanate and ammonium hydroxide, and obtained what appeared to be the same white powder which he was now able to further analyse. What Wöhler incidentally discovered was an organic compound, urea,[3] and he prepared it outside a living organism which was later deemed as a breakthrough discovery (at the time it was believed an organic compound could be obtained within living organisms only[4]).

In order to be sure of the obtained result, Wöhler should have repeated the same experiment over and over again in order to be able to finally conclude, *inductively,* that it was possible to obtain an organic compound outside a living organic system. Wöhler, however, would have considered such number of repetitions of the same experiment to be unnecessary. Why? Because he was aware that the experiment he performed was an arbitrary experiment of this type. It means that whatever happened in that experiment would happen in any other experiment performed under same relevant conditions (say, having all the glassware very clean, certain temperature or pressure maintained etc.) and with same chemicals, and no matter where and when the experiment is performed.

His inferential step was, hence, of the form: in the experiment performed, the lead cyanate could be converted into urea. There was nothing specific about the lead cyanate used, nor was the experiment performed under some unusual conditions. Hence, whatever result would be obtained, was a general one, i.e. could be generalized as holding for *any* lead cyanate. This is the inference as far as the synthesis of urea goes. If one wants to further use it to disprove vitalism, that is to defend a general claim, that an organic substance could also be obtained outside

[2] It will be seen in the following sections how this example is easily transferred to modern experimental physics.

[3] The equation of the chemical reaction in question is: $Pb(OCN)_2 + 2\ NH_3 + H_2O \rightarrow PbO + NH_4OCN \rightarrow H_2NCONH_2$. The last chemical formula is the formula for urea.

[4] Actually, the full history of the refutation of vitalism (the then prevalent doctrine that organic compounds characteristic of the living organic systems could only be obtained within such systems) is a bit more complex. For, although Wöhler did perform the very first such chemical reaction of synthesis of organic molecule from inorganic ingredients, his ingredients originally came from living substances and so, some claimed, a part of *vis vitalis* (the living force which was actually responsible for producing organic stuff) could have somehow survived and affected the whole process. Wöhler's student Hermann Kolbe is credited as the one who was able to obtain the organic substance (acetic acid which is the main ingredient of vinegar) in a wholly inorganic process from carbon disulfide (Ramberg 2015).

a living organic system, then one only needs to establish that for simillar chemical reactions (like the one Wöhler's student Kolbe performed) again there are no further relevant parameters or conditions which were not already present in Wöhler's original experiment (if it really had been perfectly designed which it was not as explained in footnote 4). Of course, one might want to test as many such reactions as possible to try to synthesize all the organic compounds from all the imaginable inorganic ingredients (which has been a larger portion of chemical research since the days of Wöhler!) but that amount of effort is wholly unnecessary in ordert to prove that at least one organic compound can be synthesized from inorganic substances and so to refute vitalism as well.

Now, we do not claim that all the results of experiments or all the discoveries in science were done by following enumerative induction, although for many one might believe that they were. What we do claim, is that all of those results that were thought of as examples of enumerative induction are actually examples of universal generalization. The art of experimentation is then chiefly consisted of finding the set of arbitrary parameters which will allow the reproduction of the phenomenon in question and not hinder its realisation, hence allowing for the relation between the right parameters to emerge for the observer. Here one can also think of further examples from physics, such as the discovery of Boyle's law (of the inverse proportionality of volume to pressure of the gas), or gas laws in general (where there is always a direct relation between two parameters). Neither Boyle, nor any other physicist involved did think there was any need for repeating the same experiment over and over again. It is true that one does repeat a certain experiment testing the dependence of certain number of parameters several times, but not because we should be more certain of the result after the n-th measurement, but because we want to minimize the errors that will, of course, always be present, nevertheless not compromising the result of the measurement.

Indeed, our analysis is not limited to physics only, although we deliberately decided to focus more on (fundamental) physics research. An example from chemistry—paradigmatic for that whole science—was already given. One can also think of many similar examples from biology. Take for instance the most fundamental discovery that every living organism has genes. Once genes were discovered in many examplars of living organisms and their function determined in any one of them, it was certain what their function will be in the specimen of the yet undiscovered species. Surely no one would doubt the degree of confidence of such a result. But can this degree ever be achieved by inductive reasoning alone?

Whenever we infer from an arbitrary situation (or object of the domain) to a general situation (or any object of the domain) we are applying the universal generalization, a deductive rule of inference. Formally we write:

$$Fa \vdash \forall x\, Fx,\ a \in D,$$

a is an arbitrary object of the domain (D), i.e. the name to be generalised upon must occur arbitrarily.

4. *How is research done in modern physics?*

To re-enforce our conclusion from previous section, that the logic of scientific research in physics (and more broadly natural science) has nothing to do with enumerative induction, we will here consider how is research done in modern physics. First, the analysis will be given due mainly to Eugene P. Wigner (1963; 1965) of the level of knowledge reached and the structure of modern physics, which should shed light on what significant changes happened already in the first half of the twentieth century fundamental physics (meaning quantum theory, relativity theories and quantum field theories). These changes were in how the theoreticians (among others Wigner himself[5]) changed the way of thinking about fundamental problems as well as the way the experimentalists changed the practice of setting up experiments. Second, we will offer what we believe should be the Wignerian reading of a class of experiments, namely the reactions between particles in particle physics.

As there are many accounts (Kaplan 1998) of scientific induction, we shall here focus on enumerative induction. However, it is our plan to undertake an expanded study of how essentially the same critique, based on Wigner's account of the structure of physics, can be used to argue against other types of inductive reasoning. One more caveat is required before proceeding further regarding the Norton's theory of material induction (e.g. 2003; 2005; 2010; 2014) as induction based on material facts that are relevant for the inductive case at hand and without relying on some universal inductive schema. We find Norton's approach very convincing in general, but feel that one can make a step further and deny that there is induction at all in science. Again, this will be elaborated in detail in a further work.

By looking at mostly physics before the twentieth century, one might be excused in thinking that (1) there is not much difference between physics and any other fairly established natural science, say chemistry; and (2) that if the logic of physical research is not always enumerative induction, it is by all means inductive logic of a kind. We, on the other hand, strongly believe, that neither (1) or (2) is acceptable. Why not? Let us look at the two cases individually. Firstly, why would (1) not be acceptable? Modern physics, since the advent of Einstein's

[5] Eugene P. Wigner was one of the first generation of quantum theorists and contributed significantly to research on quantum theory (applications of group theory to quantum mechanics) and its interpretation (especially the so called *measurement problem*) as well as to the theory (Wigner 1965) of symmetries of equations of physical laws which is what will mainly be of interest in this paper. For his contributions to fundamental research in theoretical physics he was awarded the Nobel prize for physics.

relativity and quantum theory (so since the beginning of the twentieth century) became much more general than any other science before or since. We will not go so far to state that it became akin to, say, applied mathematics, but the degree of generality of the most fundamental laws of physics as well as their great reductive power (to serve as foundation to the laws of almost all of chemistry, and therefore much of biology or geology etc.) is quite alike theories in the mathematical sciences. Furthermore, physics in general and theoretical physics in particular, employs great many mathematical techniques not only in what might be called its computational schema, but also in the way physicists think about the laws of nature. One example is the requirement that all the laws must be given as mathematical equations of a sort, most often as partial differential equations. Now, there is no such generally pronounced, and most definitely not generally accepted, view regarding, e.g. the laws of biology, or even genetics (which is much more mathematical than the average branch of biological science).

Secondly, why would (2) not be acceptable? One cannot escape the question of whether (2) is somehow not quite the best suited account of the logic of physics research, once we appreciate: (a) the crucial differences between physics and other (natural) sciences, (b) the fact that its statements possess the degree of generality that statements of no other science even remotely approach, (c) how strongly mathematical its laws are in their character, (d) the level of abstractness of theoretical physics, (e) the philosophical nature of the deepest questions physics deals with, (f) the fact that we derive laws from other more general laws (often without even doing experiments to corroborate the derived laws!), and finally, (g) how we derive whole theories within physics from a more fundamental theory, or by linking a theory to another theor

5. *Wigner's account of the structure of physics*

Wigner in (1963) and to a lesser extent, but in more detail for some of the points, in (Wigner 1965), offers a very plausible account of the whole of physics which is based on our best fundamental theories as well as our landmark experiments. In fact, it can be said his account in the meantime became the keystone of the mainstream approach to discovering new laws of physics. His interpretation of physical theory is based on a symmetry approach to the laws of physics, a movement in physics research initiated by Einstein and founded on mathematics of Hermann Minkowski, Hermann Weyl and Emmy Noether (Rosen 1983). After having discovered that in spite of physics not after all being able to give the spatio-temporal description of phenomena in absolute terms of Newtonian system, there were still some quantities and, more generally, mathematical structures, which remain unaltered when the observer's reference frame is changed—the so called *invariants,* Einstein saw this as a guide for developing new theories. He saw what was later developed as theory of invariants under symmetry

transformations as a new general framework for physics. *Symmetry* here means a transformation which preserves some structure (say a mathematical equation which expresses a law of physics) given certain change in variables (say changing the coordinates). Noether showed that to each so called *geometrical symmetry principle* there corresponds a law of conservation of a certain physical property (e.g. to a symmetry transformation with respect to spatial coordinate corresponds the law of conservation of linear momentum). It was later shown that one can (in quantum theory) make like connection for other symmetries. Given that the laws of conservation belong to the category of the most abstract and universally valid laws, one can find way in justifying Einstein's, at the time, bold claim that there is a symmetry approach to discovering the laws of nature.

Wigner stated this symmetry approach especially succinctly (and best in his Nobel prize winning lecture of 1963). In our research in physics we begin as ever with observations, more or less complex in nature or execution of experimental setup required to make these observations. At the next stage of the process of discovering laws there are certain generalizations from the observations: e.g. we abstract the specifics of the region of space and the interval of time pertaining to the observations made, or we abstract the material out of which the tested object is made etc. These *first-instance-generalizations* Wigner calls *correlations*. The correlations might be very crude and not of great degree of generality, which means that they will usually be expressed as mere approximations. Hence, valid only under certain conditions, say, Ohm's law of resistance in electric circuits is valid only for a very limited range of temperatures and materials. We can then perform further experiments to test the range of certain conditions, and here we might as well be using inductive inference techniques, but more on that will follow in the next section. So let us suppress judgement on the issue at this point. The process of further testing and refining the approximations can last for quite a time, sometimes centuries (as in the case of trying to find or refute the luminiferous ether), or for millennia (in case of discovering atoms!). The most important, however, is the next stage in development of a physical theory. And this Wigner calls the stage of forming the more general laws, which indeed can sometimes turn out to be the most general, the so called *correlations of correlations*. The laws of conservation (or, what turns out to be the same, the symmetry principles) are the most general example of correlations of correlations. We discovered each such law by the usual process of positing (hypothesizing) and experimenting on a small sample and for a limited range of values of a certain parameter. In the end, however, we have been rediscovering such regularities over and over again to the point that nowadays practically no physicist doubts the universal validity of the laws of conservation.

After we realized the general validity of the symmetry principles—and this is the crucial point in Wigner's analysis—we are better equipped for discovering further laws of physics which will be of lower level of generality and will, therefore, depend on the symmetry principles. This dependence is twofold:

1. The very existence of the lower level laws depends on the existence of higher level laws, and ultimately all the laws depend on the most general laws, some of which will be the symmetry principles.
2. The validity, or truth, of the lower level laws will depend on the validity of the higher level laws.

As Wigner himself explains regarding (1):

> It is natural, therefore, to ask for a superprinciple which is in a similar relation to the laws of nature as these are to the events. The laws of nature permit us to foresee events on the basis of the knowledge of other events; the principles of invariance should permit us to establish new correlations between events, on the basis of the knowledge of established correlations between events. This is exactly what they do. If it is established that the existence of the events A, B, C, . . . necessarily entails the occurrence of X, then the occurrence of the events A', B', C', . . . also necessarily entails X', if A', B', C', . . . and X' are obtained from A, B, C, . . . and X by one of the invariance transformations. (Wigner 1963: 10)

An example will be described in detail in the next section. It should also be noted that in the sense Wigner understood—and modern physics understands—invariance transformations (again, just another term for symmetry principles), they are to serve the purpose of a kind of selection principles, so allowing physicists to select among the several proposed possible new correlations. The one that will always be selected is the one which is in accord with symmetry principles (which usually means, one or more conservation laws). In this sense, a possible correlation cannot be declared a law of physics—so cannot really exist—if it would violate a law of conservation.

As for (2), Wigner makes the following remarks:

> The preceding two sections emphasized the inherent nature of the invariance principles as being rigorous correlations between those correlations between events which are postulated by the laws of nature. This at once points to the use of the set of invariance principles which is surely most important at present: to be a touchstone for the validity of possible laws of nature. A law of nature can be accepted as valid only if the correlations which it postulates are consistent with the accepted invariance principles. (Wigner 1963: 12)

In other words, if we need to assume the validity of the invariance principle(s) in order to accept the newly proposed law as valid (or, more cautiously, potentially valid), so to assume more general principle in order to prove that the specific, and more particulate, law holds, it means we do not have inductive reasoning at play, but at least in part also a form of deduction. Which form, remains to be examined. What

we propose is that at least in some instances of reasoning in physics, or science, it is universal generalization.

Before we proceed to examine a typical case of such reasoning, a further remark is required in order to complete the exposition of Wigner's account of physical theory and, indeed, of physics research as such. Although symmetry principles are very important in physics, it would not be all that good if everything was symmetrical at all times. Pre-requisite for even contemplating an experiment is to know (and appropriately materially realize) the so called *initial and boundary conditions*, so values of parameters which are not included within the symmetry account of the possible situation, and so present an asymmetry of a sort. Only with full specification of all the relevant symmetries, other more general laws and initial and boundary conditions might we approach discovering a new law!

6. *Experiments in particle physics and conservation laws*

The knowledge of conservation laws (symmetry principles) is of paramount importance for not only performing experiments but for even contemplating a new experiment in particle physics or nuclear physics research. What is the reason for this? It is the fact that a nuclear or, generally, a reaction between particles cannot take place unless all the relevant conservation laws are satisfied by the reaction. Physicists have, starting from around the beginning of 20th century up to today, discovered that a reaction between any number of any type of particles can in principle happen given that there is enough energy and that the specific conservation laws are satisfied. For each reaction there is the accompanying list of conservation laws[6]. For example, the list for the reaction[7] of nitrogen (^{14}N) with alpha particle (^{4}He) which has oxygen (^{17}O) and a proton (^{1}p) for products—the famous first ever nuclear transformation of elements, performed in Rutherford's team—would be:

> law of conservation of energy,
> law of conservation of momentum,
> law of conservation of angular momentum,
> law of conservation of number of baryons (this is actually easy to show from the equation of reaction, as 14 + 4 = 17 + 1),
> law of conservation of charge (the calculation is same as for the number of baryons if we assume all the particles are bare positive charges).

If any of the listed laws would be violated by what was the proposed reaction, physicists would immediately know that the reaction would not

[6] A good and standard survey of the role of conservation laws in particle physics reasearch and their connection to symmetry principles is (Henley and García 2007: 195–220).

[7] The reaction equation in standard notation is: $^{14}\mathrm{N} + {}^{4}\mathrm{He} \rightarrow {}^{17}\mathrm{O} + {}^{1}\mathrm{p}$.

take place and would not even start preparing the experimental setup. The emphasis is on the fact that there is such a complete list for each imaginable reaction and that physicists can check whether a reaction satisfies all the laws from the reaction-specific list.

Let us pause here and ask, *But how can physicists know there is such a list?* Obviously, each conservation law was discovered first as a singular fact of observation, say, it was noticed that the law of conservation of charge is valid for some chemical reactions, and later it was noted that it holds for nuclear reactions too, and so forth. Each time, however, it was valid for a particular instance of a specific reaction. The problems of inductive method of inference are already all there. Let us mention but a few:

The problem of repetition: How do we move from an observation valid for an instance of a type of experiment (a type of reaction[8]) to a conclusion valid generally for all instances of a type of experiment? Next, how do we move to establishing the same conclusion (that a particular quantity is conserved) for a different type of experiment but within the same domain of experiments (reactions between particles of certain type)?

If we take recourse to enumerative induction to make the first generalization, then the question arises, what if there appears a case of an instance of a reaction of a certain type (like the one above mentioned) for which a certain law does not appear to hold? What is the procedure then? We test again, but for what: to disclaim the negative result hitherto found, or to reconfirm this negative result, thereby in effect negating that the particular law is valid for a particular type of reaction? It is not clear, and *prima facie* cannot be clear, as we, by embracing only inductive methods of reasoning in science, cannot accept any *a priori* given fact, or any deductively posited fact. We believe the method here—and in practice of physics (or for similar situations in other sciences) might rather be universal generalization. It makes much more sense, for the reason it avoiding the aforementioned dilemma and also for it immediately being clear how to generalize not only to other instances of the same type of experiment, but also to similar types of experiments (other reactions of different particles or particle type). Taking the other instance of one type of reaction or changing for a reaction between different particles but of the same type of reaction, or switching to another type of reaction is just another arbitrary name to generalize upon.

The problem of generalization: Moreover, if the method of inferring is allowed to be from the range of deductive methods, then it is by no means unusual that we should be guided by other deductively inferred

[8] By a type of reaction it is roughly meant any reaction between a certain type of particles (e.g. a nuclear reaction is between nuclei, decay processes are transformations between nucleons, or constituents of a nucleus, etc.).

facts. Such as the fact that symmetry principles are used across the disciplines of physics, that they can guide research in physics in general (as Einstein and a battalion of first class physicists have been showing for over a hundred years now) and that there is a universally (and mathematically precisely) established connection between symmetry principles and conservation laws (Noether's famous theorems). Finally, as Wigner reasoned, we actually assume the universally valid conservation laws—and a reaction-specific list—each time we embark on testing another possible reaction between particles, or probing matter at a higher energy level, or trying to find a new particle (which is always a product in some particle reaction), most recently (in 2012) Higgs boson particle. If any of the laws on a reaction-specific list of conservation laws is violated by such a reaction, we know in advance of actually performing the reaction that it will not go.

The aprioricity of knowledge: If induction is the whole story behind reasoning in science, there really cannot be any talk of *a priori* knowledge of facts or laws, or theorems. There is always the problem of validating our inferences based on such assumptions and without deductive techniques admitted on the same footing with inductive ones. Take the last claim we made in the previous paragraph, that we can know in advance whether a reaction will go. It might seem innocent enough, indeed a practicing nuclear or particle physicist does not give it a second thought in a day-to-day laboratory work. But what a claim it is! We can know whether something will happen in advance of it happening—and we can know it with certainty, if it will or will not happen! But, making inferences by induction only, we could never reach such certainty!

Moreover, think of how we actually got to this claim: at the very first we observed a singular fact for an instance of a particular nuclear reaction; then we assumed it for all such nuclear reactions; then we generalized that a discovered correlation (a law of conservation) is valid for all reactions in nuclear physics; then we found same law holds for an instance of a reaction between some particles beyond the domain of nuclear transformations, so for a reaction in particle physics; then we generalized for all reactions in particle physics. Finally, we do not anymore question the validity of the discovered law of conservation at hand, or, for that matter, of any of the conservation laws: no one actually anymore investigates the validity of conservation laws in particle physics, they are ASSUMED, indeed so much so, that no planned experiment will ever go operational if only one of the laws from the reaction-specific list is found to be *just theoretically* violated by a reaction in question. As Wigner said, symmetry principles (or conservation laws) are to be regarded as *a touchstone for the validity of possible laws of nature.*

7. *Conclusion*

Starting with analysis of an example of a thought experiment which uses a deductive rule of inference and moving through examples from basic physics and chemistry to, finally, paradigmatic example of experiments in modern particle physics, we are drawn to conclusion that a large and significant portion of physics (science) is deductive in nature. We tried to demonstrate that what were previously thought as prime examples of application of (enumerative) induction in physics or chemistry can best be interpreted as examples of application of universal generalization inference rule. Furthermore, and by relying on an elaborate analysis of Eugene P. Wigner (one of the pioneers of quantum and nuclear physics as well as one of the foremost theoretical physicists of his generation), we showed that a deductive schema of guiding the research in physics is really the most appropriate to at least fundamental parts of that science. It is our aim to review other main purported inductive schemas and to compare with our own approach in the near future.

References

Franklin, A. and Perović, S. 2019. "Experiment in Physics." *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition). Edward N. Zalta (ed.). URL = <https://plato.stanford.edu/archives/sum2019/entries/physics-experiment/>.

Frege, G. 1884/1967. *Die Grundlagen der Arithmetik/The Foundations of Arithmetic*. Translated by J. L. Austin. 2nd rev. ed. New York: Harper and Brothers.

Galilei, G. 1638 [1914]. *Discorsi e Dimostrazioni Matematiche intorno a Due Nuove Science /Dialogue Concerning Two New Sciences*. Translated by H. Crew and A. De Salvo. New York: The Macmillan Company.

Henley, E. M. and Garcia, A. 2007. *Subatomic Physics*. Singapore: World Scientific.

Horsten, L. 2019. "Philosophy of Mathematics." *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). Edward N. Zalta (ed.). URL = <https://plato.stanford.edu/archives/spr2019/entries/philosophy-mathematics/>.

Houtappel R. M. F., van Dam H., Wigner E. P. 1965. "The Conceptual Basis and Use of Geometric Invariance Principles." *Reviews of Modern Physics* 37: 595–631.

Kaplan, M. 1998. "Induction, Epistemic Issues in." In Craig, E. (ed.). *Routledge Encyclopedia of Philosophy*. London and New York: Routledge: 745–752.

Kitcher, P. 2011. "Epistemology without History is Blind." *Erkenntnis* 75 (3): 505–524.

Lakatos, I. 1976. *Proofs and Refutations*. Cambridge: Cambridge University Press.

Norton, J. D. 2014. "A Material Dissolution of the Problem of Induction." *Synthese* 191: 671–690.

Norton, J. D. 1991. "Thought Experiments in Einstein's Work." In T. Horowitz and G. J. Massey (eds.). *Thought Experiments in Science and Philosophy*. Savage: Rowman & Littlefield Publishers: 129–148.

Norton, J. D. 2003. "A Material Theory of Induction." *Philosophy of Science* 70: 647–670.

Norton, J. D. 2005. "A Little Survey of Induction." In P. Achinstein (ed.). *Scientific Evidence: Philosophical Theories and Applications*. Baltimore: Johns Hopkins University Press: 9–34.

Norton, J. D. 2010. "There are no Universal Rules for Induction." *Philosophy of Science 77* (5): 765–777.

Polya, G. 1945. *How to Solve It. A New Aspect of Mathematical Method*. Princeton: Princeton University Press.

Putnam, H. 1979. *Philosophical Papers: Volume 1. Mathematics, Matter and Method*. Cambridge: Cambridge University Press.

Ramberg, P. 2015. "Myth 7. That Friedrich Wöhler's Synthesis of Urea in 1828 Destroyed Vitalism and Gave Rise to Organic Chemistry." In R. L. Numbers and K. Kostas (eds.). *Newton's Apple and Other Myths About Science*. Harvard: Harvard University Press: 59–66.

Rosen, J. 1983. *A Symmetry Primer for Scientists*. New York: Wiley.

Trobok, M. 2018. "The Mathematics—Natural Sciences Analogy and the Underlying Logic. The Road through Thought Experiments and Related Methods." *Croatian Journal of Philosophy* 18 (52): 23–36.

Wigner, E. P. 1972. "Events, Laws of Nature, and Invariance Principles. Nobel Lecture, December 12th 1963." In *Nobel Lectures. Physics 1963–1970*. Amsterdam: Elsevier Publishing Company: 6–18.

Croatian Journal of Philosophy
Vol. XX, No. 58, 2020

Structural Realism in Biology: A (Sympathetic) Critique

SAHOTRA SARKAR*
University of Texas at Austin, USA

Structural realism holds that ontological commitments induced by successful scientific theories should focus on the structures rather than the objects posited by the theories. Thus structural realism goes beyond the empirical adequacy criterion of traditional (or constructive) empiricism. It also attempts to avoid the problems scientific realism faces in contexts of radical theory change accompanied by discordant shifts in posited theoretical objects. Structural realism emerged in the context of attempts to interpret developments in twentieth-century physics. In a biological context, Stanford (2006) provided pre-emptive criticism. French (2011, 2012) has since attempted to answer those criticisms and extend structural realism to the biological realm. This paper argues that, though Stanford's criticism may be misplaced, and structural realism fares much better than traditional scientific realism in biological contexts, it remains a promissory note. The promise is based on shifting the focus of the debate from the status of biological laws to that of biological organization, an issue that remains a live debate within biology.

Keywords: Biology, emergentism, empiricism, holism, instrumentalism, reductionism, scientific realism; structural realism.

1. *Introduction*

Structural realism is conveniently decomposed into four related claims which form a sustained argument. Let the entities posited by a scientific theory or model[1] consist of two types: objects and structures, for

* For comments and criticisms on an earlier draft thanks are due to Steven French. This paper was begun during time spent at the Wissenschaftskolleg zu Berlin (Summer 2012). Thanks are due to the Kolleg for support.

[1] Throughout this paper, theories and models will be assumed to be entities of the same logical type, differing only in the generality of their intended domains. There is a body of philosophical literature that distinguishes between the so-called syntactic

instance, diachronic or synchronic relationships of varying complexity that hold between the posited objects. Typically, the dynamical possibilities allowed by the theory or model (what may happen over time) will be incorporated into these structures (in terms of rules governing them). The four components of structural realism are:

1. The history of science, especially in cases of radical theory change, shows that the (theoretical) objects[2] postulated even by empirically well-confirmed theories often disappear and are replaced by radically different ones—consider examples such as vortices, the caloric, phlogiston, ether, and protoplasm.
2. This aspect of scientific change critically undermines any ontological commitment to objects postulated by theories whether this commitment is only about what can be known (an epistemic claim) or about what there is (an ontic claim).
3. In contrast, some of the structures posited by theories, for instance, the laws governing the putative objects, are often resilient across radical theoretical change. The second law of thermodynamics, for instance, survived the transition from the caloric theory to classical thermodynamics and even to the kinetic theory of matter; so did many of the known chemical laws during the transition from phlogiston to oxygen.
4. Thus, in contrast to the situation with theoretically posited objects, there is ample ground for ontological commitment to the theoretically posited structures of well-confirmed theories even in the face of radical theory change.

Part (4) encapsulates the central claim of structural realism which is presumed to be a consequence of the first three parts.

The epistemic version of structural realism holds that the relevant structures comprise all that can be known; the ontic version, which is the principal locus of contemporary structural realist research, claims that these structures are all that there is (independent of any particular theories about them). Either version avoids the pitfalls of the object-oriented scientific realism that came into vogue in the 1960s and 1970s in the early post-logical empiricist philosophy of science following a general (and, perhaps, misguided) rejection of the instrumentalism associated with most of the logical empiricist canon. Structural realism also goes beyond traditional empiricism in denying incorrigible phenomenal content as the epistemic foundation for scientific knowledge and, especially, by not accepting a criterion empirical adequacy as the

and semantic interpretations of theories, with "models" supposed to be related to the latter; however, neither the goals of that project, nor the many problems with such accounts, are relevant to the issues treated in this paper. Most importantly, the usage here follows standard scientific usage in biology (and elsewhere)—see Frigg and Hartmann (2006).

[2] The term "object" must be construed broadly to include any non-relational entity (particle, field, cell, information, community, carrying capacity, *etc.*).

sole desideratum for the adjudication of theoretical commitment, the latter position being closely associated with constructive empiricism (van Fraassen 1980).

Historically, structural realism was developed with the goal of providing a viable realist interpretation of modern (twentieth-century) physics taking into account the profound conceptual changes induced by quantum mechanics as well as the special and general theories of relativity. Stanford (2006) pre-emptively criticized its applicability to biology as part of a general critique of realism about science. French (2011, 2012) attempted to answer those criticisms and extend structural realism to the biological domain. This attempt is usefully analyzed into two separate theses: (i) a critique of object-oriented realism about biology; and (ii) a tentative defense of realism about structures interpreted as biological laws[3] which, though admittedly lacking universality, apparently remain resilient under many theoretical changes.

The purpose of the present paper is to offer a critical assessment of structural realism in biology. Because structural and any other forms of realism are easy to criticize on purely philosophical grounds, especially when divorced from the practice of science, and when no alternative need be provided, Section 2 will sketch a set of positive theses that are supposed to criticize structural realism and fare somewhat better at interpreting contemporary biology. As a consequence of that discussion, Section 3 will largely endorse French's skepticism about object-oriented realism in biology but emphasize several subtleties that dilute the impact of his critique. However, and more importantly, it will also extend this skepticism to the biological laws favored by French in his defense of structural realism in biology. Section 4 will turn to the role of organization—and that sense of structure—in the history of biology, and in contemporary biology, and argue that this is where structural realism is most plausible in biology. Section 5 will question whether, even in its most plausible domain, prospects for structural realism in biology are better than dim. It will be inconclusive. Ostensibly to compensate for that, Section 6 will draw some conclusions.

2. *Positive Agenda*

It will be instructive to begin with the putatively central insight of structural realism: that certain structures (for instance, relationships between putative objects) persists over radical theory change, radical in the sense that the objects postulated by the earlier theory do not survive the same transformation. This point will be illustrated in this section using an example that instrumentalist critics of structural realism have deployed in favor of their own position, *viz*., Galton's biometrical

[3] It is open to question whether structures should necessarily be interpreted as laws or even as relationships between (adequately individualized) objects. I follow French on this point *for biological contexts*. Nothing in French's discussion—or mine—restricts structures to laws or relations.

Law of Ancestral Heredity.[4] The discussion here will bring that use of this law into question.

However, before turning to the details of that example, it is worth emphasizing (with French [2011, 201]), the general non-persistence of theoretical objects in biology. Take perhaps the single most important such object of twentieth-century biology: the gene. Two points, both of which deserve much further elaboration than will be possible here, are of relevance: (i) It is far from clear that "gene" continues to play *any* theoretical role, rather than an informal heuristic one, in contemporary postgenomic accounts of heredity.[5] Arguably, in explicit theoretical discussions of DNA behavior during cell reproduction and differentiation, the concept of a gene has no more a cognitive role in contemporary biology that what the concept of an electron orbiting a nucleus has in contemporary chemistry. If this is correct, even though much of the insights of classical genetics, in the forms of rules of transmission and expression of traits, continue to remain relevant, during the last two decades the gene has lost its the pre-eminent ontological status that it had in biology for almost a century (Keller 2002b). (ii) To the extent that certain DNA sequences can still be usefully characterized as traditional genes (most importantly, some of that tiny fraction of DNA in most eukaryotes that uniquely specify amino acid sequences of proteins[6]), these objects share as few properties with Johannsen's (1905) original "genes" as today's atoms do with Dalton's creation. For instance, thanks to ubiquitous alternative splicing, a single gene may often specify more than one phenotype (at least at the protein level). Stein (1989) has aptly pointed out that to assume the "reality" of the atom and not, say, of the ether on the basis of the persistence of one term and not of the other is no more than a surrender to the vagaries of changes in linguistic usage. The same point can be made about the persistence of "gene"; an even stronger case can be made against another pillar of mid-twentieth century molecular biology: biological "information" (Sarkar 1996).

In contrast, turn now to a discarded tradition in the study of heredity that once held considerable promise: Galton, Weldon, and Pearson's science of biometry.[7] Galton posited the existence of a "stirp" in the

[4] This example is important in this context because it forms part of Stanford's (2006) critique of structural realism which will be discussed later in the text. For a more detailed philosophical analysis, see Sarkar (1998, Chapter 5).

[5] See Perini (2011) for a good discussion and an entry into the extensive literature.

[6] The qualification "some of" is necessary to exclude overlapping genes, *etc*.—see Sarkar (1996) on this point; the qualification "uniquely" similarly avoids problems associated with alternative splicing. The qualification "traditional" is necessary because it is not at all unusual to refer to any functional DNA segment as a gene (*e.g*., Lynch [2007], Koonin [2011], *etc*.), no matter whether it is transcribed and translated, transcribed but not translated, or even plays a regulatory role in some other way—this is the heuristic or informal notion of a gene noted earlier.

[7] The best summary is Pearson (1900); Provine (1972) and Sarkar (1998) provide historical and philosophical discussion.

germinal cells of organisms which mediated the inheritance of traits from parent to offspring. On the basis of this model of inheritance, he postulated several nomological claims, the most famous of which was the quantitative Law of Ancestral Heredity which, after subsequent clarification and reformulation by Pearson, states (roughly) that the ancestral contribution to any hereditary trait of an individual organism decreases in a geometric series with distance up the family tree.[8] The biometricians were (correctly) adamant that a wealth of quantitative empirical data on continuously varying traits from the 1880s and 1890s supported the Law of Ancestral Heredity.

It is uncontroversial that the theoretical claims of biometry—in particular, Galton's stirp model of inheritance (to the extent it should even be taken to be part of the science of biometry)—were superseded and replaced by Mendel's model of inheritance shortly after Mendel's work was recovered around 1900, and after an acrimonious dispute between adherents of the two sides with the Mendelians represented primarily by Bateson but with support from others including Punnett.[9] Yet, as Olby (1966, 1987) and others have periodically pointed out, the mathematical relationship incorporated in the Law of Ancestral Heredity, interpreted as a correlation between traits of an organism and its ancestors (rather than as a "contribution" from ancestors), continues to hold in a Mendelian[10] context.

This would seem to be grist for the structural realist's mill. Stanford (2006: 182), however, is dismissive; according to him, what Olby's observation (and others that are similar) show is that:

> "the formal relationship described by the Ancestral Law [*sic*] can certainly be unearthed by sufficiently persistent digging into the corners of the theoretical description of the world given to us by contemporary genetics.
> But it is equally true that contemporary genetics does not recognize the fractional relationships expressed in Galton's Ancestral Law as describing any fundamental or even particularly significant aspect of the mathematical structure of inheritance."

Stanford continues with a dismissal of Worrall's (1989) version of structural realism.

Though the neutrality between realism and instrumentalism that I generally endorse shares some of Stanford's skepticism about realism, his dismissal of the Law of Ancestral Heredity is unwarranted. Any serious history of the Law of Ancestral Heredity must pay more attention to the pertinent detail. Pearson's reformulation of the Law of

[8] Galton's (1965) first rudimentary statement occurs in the work taken to be the origin of eugenics, "Hereditary Talent and Character"; Pearson's final statement appears in the second edition of *The Grammar of Science* (Pearson 1900).

[9] This has been extensively documented by Provine (1971).

[10] The term "Mendelian" instead of "Mendel's" is being used to distinguish between what became part of the new (Mendelian) genetics between 1900 and 1920 and Mendel's own statements which required considerable modification during the formulation and establishment of what came to be called Mendelian genetics.

Ancestral Heredity involved two related crucial philosophical moves: (i) He dropped the stirp model altogether and eschewed causal talk (of "contribution") in favor of correlation between traits. So, whether or not Galton's stirp model of inheritance (which constitutes an object-oriented ontology) is correct becomes irrelevant to the status of the Law. (ii) In general, Pearson insisted that, in the historical context in which biometry was attempting to construct a quantitative theory of evolution by natural selection, the laws of heredity should remain what will be called *phenomenological*. This move to phenomenological characterization was a consequence of Pearson's quite sophisticated positivism—but that is a story for some other occasion.

By "phenomenological" here I mean laws that employ the same (or very similar) conceptual resources as those deployed to report the results of experiments. This is a matter of degree. Some claims are more phenomenological than others; in that sense they are less theoretical than those others. Note that there is no claim here of any hard observational-theoretical distinction. How experimental reports are formulated depends on what theories are taken to be sufficiently well-established so as not to be challenged by the experiments being performed. What is at stake here is that, in the given context of research, phenomenological resources can be used to formulate claims that can be used to adjudicate between the theories that are in play. Returning to the example at hand, the Law of Ancestral Heredity, interpreted phenomenologically, could potentially be used to distinguish between Galton's and Mendel's models of inheritance. Historically, it turned out to be the case that it is consistent with both in the sense that both models *semi-formally* predict it, where mathematical predictions are deemed to be "semi-formal" if they require idealizations or incorrigible approximations.[11]

What is more important in this context is that the Law of Ancestral Heredity was taken to be sufficiently empirically well-supported to impose constraints (adequacy conditions) on permissible theorizing about heredity in the 1900–1920 period: any adequate theory of heredity had to incorporate that Law. This is seen, in particular, by Pearson's (1904a, b) own attempts to derive the law from Mendel's rules as well as Doncaster's (1910) review of recent work in heredity which discussed both Mendel's rules and the Law.[12] When Fisher (1918) began his ambitious project of using Mendel's rules to account for inheritance patterns of continuously varying traits—what led to the subsequent discipline of quantitative genetics—it was still perceived to be critical to establish consistency between the Law of Ancestral Heredity and Mendelian

[11] Here "incorrigible" means that there is no known procedure to weaken the relevant approximation—for a discussion, see Sarkar (1998: 49).

[12] Even by 1920 Doncaster had not changed his mind—see Lock and Doncaster (1920).

rules[13], hardly something to be dismissed as "persistent digging into the corners of the theoretical description of the world given to us by contemporary genetics." What Fisher showed was remarkable: the Law of Ancestral Heredity could be semi-derived from Mendelian rules.[14] An entire section of "The Correlation between Relatives on the Supposition of Mendelian Inheritance" (§ 17) was devoted to deriving that Law from Mendelian assumptions. It amounted to a reduction of the Law of Ancestral Heredity to Mendelian genetics. In fact, what Fisher achieved was the reduction of all the more salient nomological claims of biometry to a Mendelian basis. This included, for example, the rule that quantitative traits follow the normal distribution in large populations.[15] After Fisher's derivation the empirical status of the Law of Ancestral Heredity was no longer in question: evidence for Mendelian genetics was *ipso facto* evidence for that Law (at least informally).[16] What changed was that all the biometrical generalizations proved to be of decreasing utility in practical contexts of quantitative genetics, the most important ones being those of agriculture and animal breeding.

Structural realists will interpret this situation as indicating that though there should be no ontological commitment to theoretical objects (Galton's stirp of Mendelian genes), there are grounds for such commitment to the relevant structures, that is, associated laws such as the Law of Ancestral Heredity. This position can be bolstered using a wealth of examples from the physical sciences including, as noted earlier, the persistence of the second law of thermodynamics in the transition from the caloric theory to classical thermodynamics (incorporating the first law, or conservation of energy). Contrary to Stanford (2006), such an interpretation of the significance of the persistence of the Law of Ancestral Heredity is hardly far-fetched.

What skeptics of structural realism must do is to provide a more *scientifically* compelling interpretation of these developments (in the sense of a more plausible interpretation of history and practice in the relevant scientific episode). What follows is a sketch such a position, one which is supposed to provide a contrast to structural realism but does not endorse any form of anti-realism (including constructive empiricism). Rather, partly following and extending the discussions of Nagel (1961) and Stein (1989), it sees no essential difference between

[13] Fisher was neither the first nor the only geneticist to acknowledge this requirement: Yule (1902) and Weinberg (1909a, b; 1910) were among those who preceded him—Stern (1965) provides an illuminating discussion of these developments.

[14] For critical discussion, see the commentary by Moran and Smith (1966) and the discussion in Sarkar (1998, 106 –107).

[15] This aspect of the creation of quantitative genetics is discussed in more detail by Sarkar (1998, Chapter 5). See, also, Frogatt and Nevin (1971). But much more philosophical analysis would be welcome—and would not go unnoticed.

[16] In general, evidence for a reducing theory is indirect evidence for the one that is reduced (Sarkar 1998).

a sophisticated instrumentalism and a modest version of structural realism which is closer to the epistemic rather than ontic version. Ultimately, the force of the critique of structural realism being developed here should be taken to rest partly on the plausibility of this alternative view.

It will serve to present this alternative position as being constituted by four distinct substantive points followed by one polemical one which is of less importance:

1. With structural realism, it agrees that the history of science makes it impossible to defend any ontological commitment to theoretical objects (object-oriented realism).
2. Again with structural realism, it agrees that the certain structures are more resilient across theoretical change than objects. In the biological context these structures include (but are not limited to) laws though not all laws have the required degree of resilience.
3. Unlike structural realism, the resilience of these laws is explained by their phenomenological status in the context in which they are introduced or used to adjudicate between rival theories. This is a central tenet of the position being advocated here and some elaboration seems in order. In a given historical context, phenomenological laws are supposed to be theoretically neutral in the sense that the theories being adjudicated do not differ in their predictions (or otherwise) with respect these laws. By and large—and this a claim subject to historical test—in further development of a field, laws that were deemed phenomenological in one context will remain so in future contexts because it seems implausible that they will become "theory laden" with newer, typically more abstract, theoretical assumptions.[17] Thus, phenomenological laws form part of what each successive theory must explain. Consequently, they are often resilient over theory change.
4. Nevertheless, phenomenological laws need not have indefinite tenure. For instance, radical theoretical—or even experimental—change may show that the degree or type of approximation involved in accepting a phenomenological laws may make it contextually no longer admissible to deem such a law as (approximately) correct. In Section 3 it will be argued that this is, indeed, the situation of the Law of Ancestral Heredity in the light of postgenomic developments. There is no evidence in biology that there is convergence to any set of phenomenological laws that appear so safe from future rejection (or, at least, radical revision) to warrant deep ontological commitment. Indeed, if structural realism is committed to such laws as the only relevant structures, it will not fare better than object-oriented realism.

[17] Note the qualification, "by and large"—this is not being presented as an exceptionless claim.

5. The final point is polemical and historical—the cogency of the arguments presented here does not depend on its validity but it help show what, at least partly, motivates this position. Points 3 and 4 have much in common with logical empiricism, in particular, the views of Neurath, Reichenbach, and Nagel. What are being called phenomenological laws here are generalizations of what the logical empiricists called protocol sentences expressed in a physical language (the generalization being that these phenomenological laws are universally quantified over the relevant domain). Like protocol sentences, these laws are corrigible though, unlike protocol sentences, they are not the sole epistemic basis for the relevant theoretical models.[18] The attitude towards ontological commitment expressed here is also similar to that of those logical empiricists who endorsed some form of "realism" but saw it as being consistent with their empiricism in contrast to the types of realism associated with object-oriented or structural realism.

The scope of this alternative position is at present intended to be limited to biological contexts in which there are no known "deep" structures (such as symmetry groups in some physical contexts) which cannot be easily interpreted as phenomenological laws.

3. *Biological Laws and Structural Realism*

As noted earlier, a case against object-oriented realism in biology could have made profitable use of examples such as the gene or information. Equally apt ecological examples would include carrying capacity, climax community, and intrinsic growth rate. In developmental biology terms that have undergone radical shifts of empirical significance include "genotype" and "norm of reaction" (Sarkar 1999). However, the only published defenses of structural realism in biology (French 2011, 2012) rely on Dupré and O'Malley's (2007, 2009) critique of biological individuality as delimiting a unique set of (biological) objects. There are two pitfalls with this line of argument:

(1) Dupré and O'Malley's concerns are synchronic, to deny *at this time* the possibility of a unique ontology of well-defined objects constituting the biological realm. Instead, they opt for pluralism and what is called a "promiscuous realism" (Dupré 1996) about objects. Leaving aside a discussion of the plausibility of promiscuity for some other occasion, in this context what is at stake is the diachronic identity of objects across theory change because that is what structural realism denies. The problems raised by Dupré and O'Malley are tangential to this issue.

[18] Rather, they are the *explanans* in Nagel-type models of reduction (see Nagel [1961]).

(2) The second problem is both philosophically and biologically more important. Long ago, Nagel (1951, 1952) pointed out that any mereological decomposition of an object requires theoretical assumptions. Objects do not simply exist in a categorical spatial hierarchy; rather, to say that a given object consists of a specified set of parts is to make a theoretical claim, one choice among others about how to decompose a whole into its parts. The cogency of a decomposition depends on the empirical success of this theory along with the relevant theoretical claims about the behaviors of the whole and the parts (including their interactions). While Nagel made this perceptive observation in an explicitly biological context, it is relevant to all scientific contexts in which hierarchical organization is presumed. The biological context introduces an added complexity: the wholes, as well as the parts, are themselves (a) historically evolved objects[19] that (b) must be individuated using theoretical criteria—that is, beyond Nagel, even what the whole is requires theoretical specification. Physical individuals need not be organismic individuals: in most physical mammal bodies the vast majority of cells are not those of the mammalian individual *qua* mammal. (Consider, for instance, the human skin or intestine—there are 10 times as many non-human cells in the latter as there are human cells in a typical body [roughly 10^{14} of the latter].) Genotypic individuals need not be physical individuals, *e.g.*, in the cases of dandelions or aphids. In fact, what Dupré and O'Malley's (2007, 2009) analyses show is the ubiquity of the individuation problem in the metagenomic context (which is not unexpected).

To make a case against object-oriented realism on the basis of problems of biological individuality will require (i) the specification of a theoretical individuality criterion (genotypic, immunological, organismic, *etc.*) and (ii) a demonstration of the diachronic ephemerality of these individuals across theory change. French does not do this, and it remains an open question whether biological individuals, however defined (so long as these definitions are exact and explicit), are as ephemeral as, say, genes or information.

The last paragraph may well have been a digression from the argument of this paper since it agrees with structural realists that an ontology of biological objects is far too unstable to warrant "realism." What is more problematic for French's argument is the question of the resilience of biological laws. It will be instructive to return to the Law of Ancestral Heredity. It was pointed out in Section 2 that the fundamental mathematical (read "structural") claim of that law, that is, the geometric regression of correlation with ancestral relatives, survived the transition from biometry to (Mendelian) quantitative genetics.

[19] See, in this context, Buss (1987) and the commentary by Falk and Sarkar (1992).

The potential trouble is that the postgenomic era is witnessing a much more radical shift in the understanding of heredity than the shift from biometry to Mendelism (though the ongoing shift is as yet poorly understood even within biology, let alone in the philosophy of science). It was noted in Section 2 that few DNA sequences exhibit Mendelian patterns of inheritance. Add to this (i) that horizontal DNA transfer across lineages has been ubiquitous in early evolution (which, either in the number of years or in the number of generations, has been the longest period of evolution), (ii) large DNA sequences often duplicate during reproduction (and this process is now widely recognized as being critical for the generation of evolutionary novelty), and (iii) genomes tend to expand through a variety of molecular mechanisms due purely to the physics of DNA interactions (Lynch 2007). It is questionable that the Mendelian rules will survive this transition except as approximations applicable to a tiny fraction of inherited traits (though these are the ones that dominated research in twentieth-century biology because the Mendelian rules they followed made them easily tractable). It appears unlikely—though this is as yet unproven—that the Law of Ancestral Heredity will survive this ongoing transition any better; worse, given that it is an approximation even in a Mendelian context, it will become irrelevant. The philosophically salient point is that even phenomenological laws do not have indefinite tenure though they generally have longer ones than theoretical objects.

A potentially more interesting "law" is the Price equation on which French (2012) aptly focuses. This equation, which has recently been the focus of sustained interest within evolutionary biology (Frank 2007), began its remarkable career as an intended reformulation of what Fisher (1930) called the fundamental theorem of natural selection (Price 1972). However, it turned out to be more general in two important ways: (i) it can recursively incorporate the operation of selection at multiple levels of a hierarchy, and (ii) it does not depend on the details of the Mendelian model of inheritance. This generality makes the Price equation more akin to a constitutive framework in which a variety of laws can be formulated (or, equivalently, models can be constructed) than to an individual law—this point will be relevant in Section 6.

But there are ample grounds at least for caution, perhaps downright skepticism. While Fisher regarded his theorem as fundamental, and a minority cadre of very vocal theoretical population geneticists have followed him in extolling its virtues, it should not be forgotten that the other two major founders of theoretical population genetics, Haldane (1932) and Wright (1930), were skeptical of its significance (Edwards 1994). If taken as an exact claim, that is, its mathematical form is supposed to capture the operation of selection *in toto*, the assumptions of the theorem hold for vanishingly few cases. The same problem carries over to the Price equation: more technically, both Fisher's theorem and the Price equation make strong and debilitating assumptions of the

additivity of the effects of alleles (or their equivalents at other levels of organization).[20] Now, if Fisher's theorem and the Price equations are taken to be approximate, then it is less than clear what ontological significance should be attached to the persistence of such a structure. (However, both the theorem and the equation now become applicable to many more situations.) A way out would be to regard either of them as an idealization but then it would be one requiring a host of counterfactual assumptions: it is up to structural realists to show how such extreme idealizations can ground deep ontological commitments. This may not be an impossible task. Meanwhile, at present, there is ample ground to doubt the significance of the Price equation—moreover, and perhaps most importantly, what will happen to it in post-genomic accounts of heredity also remains far from clear. It takes a lot of faith to assume it will be resilient in the way that structural realism requires. Worse, no other putative biological law provides better prospects for structural realism.

4. *Biological Organization and Structural Realism*

The failure of biological laws to underpin structural realism does not sound the death knell of that doctrine in the biological context. Rather, structural realists would do well to focus their attention on biological organization. This means a shift of focus from what may be called nomological particulars (individual laws) to constitutive frameworks in which these nomological claims can be formulated.[21]

Historically, two distinct themes have been important:

(1) Since the late eighteenth century, and even after the demise of traditional vitalism in the nineteenth century, biology has persistently accommodated research programs based on the assumption that biological organisms have some feature(s) that distinguish them from what may be called purely physical (or chemical) structures. In general, there was no claim that biological organisms exhibited mechanisms at variance with the known physical (and chemical) ones; rather, invoking only these mechanisms was deemed insufficient for the satisfactory explanation of biological phenomena. The various research programs that incorporate such assumptions may be distinguished into two groups[22]:

[20] See, however, Frank (1997) who defends the additivity assumption but nevertheless accepts that it imposes some restrictions.

[21] That is, within a constitutive framework, a variety of laws can be formulated. Typically, in biological contexts these laws are called models.

[22] The characterizations given here intentionally avoid the issue of reductionism which will be fully broached in Section 5.

i. Teleological holism, discussed in Section 4.1, which emphasizes function and teleology in a way that was supposed to subordinate the relevant explanatory behaviors of parts to goals that were only specifiable by reference to the whole.

ii. Structural emergentism, discussed in Section 4.2, which emphasizes what is typically referred to as the emergence of systemic properties which are supposed to be at variance with the properties of the constituent parts of these systems.

(2) Since the nineteenth century there also has been a long—and, at least arguably, so far futile—search for laws of form: principles of structural organization which are supposed to explain what Raff (1996) called the "shape of life." These laws of form are supposed to explain why, for instance, all animal embryos at an early stage of development have either two-fold or five-fold symmetry (and no other). The salient research programs will be discussed, though only very briefly, in Section 4.3.

In the present context, what is relevant is that these research programs emphasize structure over objects. In the case of developmental form, the structure is clearly spatial; in the cases of teleological holism and structural holism, the structure may be embedded in an abstract space but may also be spatial in nature, as is usually the case for the structural emergenticists. The details that follow are intended to show why these programs *may* support structural realism.

4.1. *Teleological Holism*

An epistemological characterization of the assumptions of research programs subsumed under this category is relatively straightforward. Organisms (or other wholes) are supposed to be categorically described as having goals. Here "categorically described" means a type of description that is necessary to understand these organisms (wholes) *qua* organisms (wholes). In some form or other, such a view of living organisms can be traced back to Aristotle; it is a plausible (and was a popular) interpretation of the second part of Kant's *Critique of Judgment*. With Kant, nineteenth-century teleological holists such as von Baer generally held that the mechanisms operating within living organisms were no different than those also operating in non-living matter.[23] However, to explain living phenomena satisfactorily required reference to the goals of the whole: why a part does what it does depends on its structural relationships with other parts with which it forms a whole; these relationships establish its functional contributions to the goals of the whole. Consequently, any determination of the set of

[23] A complex history is being selectively summarized—and perhaps caricatured—here for philosophical purposes, possibly to the extent of parody. See Lenoir (1989) for more detail.

mechanisms that are explanatorily relevant to the living phenomena that are to be explained must take into account how the parts are structured so as to comprise the whole. As Lenoir (1989: ix) puts it: this was "a period in the history of the life sciences when the imputation of purposiveness was not regarded an embarrassment but rather an accepted fact, and when the principal goal was to reap the benefits of mechanistic explanation by finding the means of incorporating them within the guidelines of a teleological framework." A more radical version of these claims would go further to argue that what the parts are is relatively irrelevant compared to the structure: this is the form that teleological holism took under the guise of cybernetic models in the mid-twentieth century (see below). An ontological characterization of these doctrines adds an ontological gloss on the claims of this paragraph (but does not change any other feature).

The mid- and late nineteenth century saw the relentless progress of mechanistic explanations in the life sciences, that is, explanations of the properties of wholes from those of their constituent parts and their interactions (Sarkar 1998). Nevertheless, a form of teleological holism became fashionable in physiology through sustained advocacy by Bernard (1865) and his insistence that the physiological behavior of parts of an organism could only be understood in terms of the context in which these behaviors occurred, the context being specified by the other parts of the functional whole. Other physiologists including Christian Bohr and J. S. Haldane in the early twentieth century explicitly embraced similar doctrines.[24] The critical assumptions were (i) that physiology was intrinsically about function and (ii) that function could only be understood by subordinating the behaviors of parts to that of the whole. Of particular interest were co-operative phenomena, in which the increase in the number of units results in a non-linear increase of effect, for instance, the S-shaped association curve between hemoglobin and oxygen that Bohr established (the "Bohr effect"). These cases often displayed feedback regulation—a drop in the response after saturation with oxygen, a feature seen in the S-shape of the hemoglobin-oxygen association curve. Structurally what mattered is how the system was constructed together and how the parts with their functions interacted with each other. In this sense these views were very similar to those of teleological holists of the nineteenth century (and, by and large the physiologists were explicit in admitting the influence of Kant's third *Critique*).[25] The term "holism" was coined later by Smuts (1926), though mainly in an evolutionary context, to embrace these views.

Meanwhile, the emergence of biochemistry as an organized discipline under G. W. Hopkins in the 1920s and its empirical successes saw

[24] See J. S. Haldane (1906, 1914); on Bohr, see Tigerstedt (2012). Their views also had an influence of the non-mechanistic theses promoted by their more famous offspring: Niels Bohr and J. B. S. Haldane (Holton 1970; Sarkar 1992b).

[25] See, for example, J. S. Haldane (1914).

mechanistic explanation return to the forefront in contexts in which holistic physiology had once reigned unchallenged (Sarkar 1992a). However, models of feedback regulation, beginning in the mid-1950s, typically based on Wiener's (1948) cybernetics, gave teleological holism a new lease of life.[26] The self-regulation of enzyme (more specifically, lactase) production in bacteria (*Escherichia coli* in this case) in the presence of the relevant substrate (in this case, lactose) emerged as a problem of experimental investigation[27]—the result was the operon model, the significance of which will be further discussed in Section 5. Monod (1971) later dubbed this work as "molecular cybernetics."[28] Suffice it to note that the interest in the regulation in biological systems has had a continuous history since the 1950s resulting in the current emphasis of gene regulatory networks (GRNs); some of these developments will be taken up in more detail in Section 4.2.

In the present context what is most salient is the extent to which such models of self-regulation make specific (that is, detailed) and general (that is, applicable to a wide variety of cases) assumptions about structural organization. Historically, at the very least, these models universally assumed a network structure of interactions between the unit components constituting the whole (that is, the interactions could not be reduced to a single chain), and typically assumed loops (enabling feedback). It can be assumed without loss of generality that the mathematical structure required by these models is that of a directed multigraph[29] (which, for ease of formal analysis, is typically reduced to a directed graph). The structural realist thesis is now straightforward to state: the edge sets, that, is their topological or connectivity features (what types of connections there are), will show resilience across theory change even when the identity of the vertex set changes. If so, in such models, the explanatory weight (however that is explicated) is borne by the structure rather than the objects—as structural realism would require. There will be more on networks and multigraphs in Section 4.2 below.

4.2. *Structural Emergentism*

The focus will continue to be on networks modeled as directed multigraphs. However, there is a critical difference between the research programs considered here and those mentioned at the end of the Section 4.1. The models analyzed here do not insist on some special role

[26] The importance of cybernetics to mid-twentieth-century science is hard to understand today (because of its apparently total failure) but can hardly be overstated—see Heims (1991).

[27] Schaffner (1974) provides a detailed history.

[28] Sarkar (1996) provides background.

[29] These differ from ordinary graphs insofar as edges and vertices can be of more than one type; thus, for instance, more than one edge (each of a different type) can join two vertices.

played by the goals of function of the whole or on whether explanations using constituent parts must refer to the wholes. There is no explicit teleology in these models. Instead, most (though not all) such models are concerned with whether the topology of the edge sets are more strongly implicated (that is, bear the most explanatory weight) in the behaviors of networks as systems compared to the vertices (objects) of the multigraph. If so, in this mitigated sense, the behaviors of systems involve "emergent" properties, dependent on how a system is put together rather than of what it is made.[30] An example, discussed in some detail by Sarkar (1998: 168 –173), is the molecular explanation of dominance which was an ubiquitous feature of classical genetics: why, for some traits, the heterozygote is phenotypically identical to one of the homozygotes. The best explanation so far seems to be in terms of the topology of the reaction networks connecting the DNA specifying the alleles to the molecular structures corresponding to the phenotype. (However, experimentally, the issue is far from settled.)

The relevance of such a situation to structural realism is straightforward: in cases where structure matters more than identity of the parts, it is highly likely that the topology of the network will be resilient across many theory changes involving revisions of the identity of the units (that is, the edge sets will be more resilient than the vertex sets of the multigraph). Moreover, and this point deserves emphasis, such resilience is logically independent of whether there is any more interesting sense in which the networks exhibit emergent behavior. Thus, though this section is on structural emergentism, the emphasis is on structure rather than on emergence. In what follows, to focus on structure, the issue of emergence will be intentionally ignored.

Complex networks of this type constitute the central metaphor of the apparently emerging discipline of systems biology that has become a component of postgenomics. Such complex networks are also supposed to explain ecological behavior—in particular, the emergence of large-scale order—over both large spatial and temporal scales.[31] Most models of "complex adaptive systems""—yet another popular metaphor of contemporary science—are network models. In fact, to the extent that an alleged science of complexity exists (and there is room for skepticism on this point [Horgan 1995]), it is a science of networks. The relevance to immunology of network models—under the rubric of idiotypic networks (Jerne 1974)—has long been postulated though never fully satisfactorily demonstrated.

Turning to only somewhat less speculative areas, complex gene regulatory networks (GRNs) are supposed to provide, at present, the most viable candidates for understanding organismic developmental

[30] This is intended to be a minimalist and neutral epistemological characterization of emergence. For an introduction to the tendentious philosophical disputes regarding this doctrine, see Bedau and Humphreys (2008).

[31] See, for example, Pascual and Dunne (2006) and Fortuna (2007).

cycles (from germinal cell through the adult stage to reproduction) (Davidson 2006). In this field, they have an illustrious pedigree, going back to Boveri's work at the beginning of the twentieth century, and continuing to what is called developmental evolution today.[32] Current GRN models can be traced back to Britten and Davidson's (1969) model which was the first putative general model of eukaryotic gene regulation given the complexities of eukaryotic genome structure that had begun to be recognized in the 1960s. Though there was some formal similarity between this model and the earlier operon model (for prokaryotic gene regulation—see Section 4.1), and textbooks of the period routinely (over)emphasize this aspect[33], unlike the operon model, the Britten-Davidson model was not concerned at all to explain feedback regulation; rather its aim is to explain tissue differentiation and the development of complex form—hence its inclusion in this section rather than in Section 4.1. Though largely ignored for a generation, as the complexity of eukaryotic genetics seemed to defy any modeling strategy (Sarkar 1996), a much-modified Britten-Davidson model and its descendants, in the form of GRNs, have returned to the forefront of research in cell differentiation and organismic development in postgenomics. Whether these models live up to the hopes of their enthusiasts remains to be seen—let me note that, among biologists, there remains ample ground for skepticism.[34]

From the perspective of this paper, these developments suggest the following conclusion: to the extent that the biological sciences may have any universal mathematical structure (that may potentially play the same unifying role as symmetry groups play in modern physics), that structure seems to be that of directed multigraphs. Perhaps what structural realism should focus on is on the demarcation of the types of directed multigraphs that are relevant for biological theory, and then a classification of these based on the roles they play in various biological sub-disciplines.

4.3. *Developmental Form*

The final set of research programs to be considered here consists of models that have remained speculative throughout their roughly 150-year history. These are macroscopic models ("macroscopic" in the sense that they are concerned with large spatial structures) of developmental form, how organisms produce their adult forms through the history of interactions between the physical contents of germinal cells and their environments. One important class of such models consist of those that rely on details of the physical interactions of the molecular constituents—perhaps the best-known such model was that introduced by

[32] Thanks are due to Manfred Laubichler (unpublished work) for providing this history.

[33] See, for example, Lewin (1974).

[34] See Newman (2019).

Turing (1952), based on equations for reaction-diffusion systems, and capable of generating a wide variety of spatial forms.[35] However, what are most pertinent to the question of the plausibility of structural realism are models that are based on spatial regularities and transformations that are independent of assumptions of the details of the underlying physical interactions. Nineteenth-century morphologists such as Cuvier established several such rules across phylogenetically related sets of taxa in a period when virtually nothing was known about the underlying physical or chemical mechanisms. Embryologists followed their lead by producing similar analyses not only on adult forms but on the developmental stages of organisms generating interesting possibilities, for instance, the hypothesis of the existence of a near-universal phylotypic stage for many animal phyla (Raff 1996).

In the twentieth century, D'Arcy Thompson's (1917) *On Growth and Form* provided a remarkable compendium of mathematical rules that transform spatial features of one taxon to phylogenetically related spatial features of others. Thompson's project involved a shift away from evolution (and history) to questions of form and universal rules that may govern their genesis. For structural realism, what is intriguing is that such mathematical transformation rules would likely be independent of the details of the underlying physical (or chemical) basis and thus be resilient to changes of the ontology of the objects being postulated by models of development. Since the 1980s, with the advent of ubiquitous computation, a further set of models for spatial patterns have been investigated, especially using cellular automata: these models show how very simple generative rules may lead to complex spatial patterns. Beyond organismic development, these rules may also be relevant to the appearance of long-range spatial and temporal patterns in ecology (Ermentrout and Edelstein-Keshet 1993).

Returning to organismic development, what remains unclear, is the *nomological* status of D'Arcy Thompson's and similar transformation rules, whether they are any more than piecemeal (accidental) generalizations that reflect no deep structure of developmental processes. Skeptics of laws of developmental form have ample ammunition on their side: after at least 150 years there is no clear example of a single well-established theoretical law of form. However, whatever be the merit of these hopes, the quest for laws of form seems to continue to find deep resonance in the intuitions of many developmental biologists.[36] Arguably, it is even part of what motivates the recent excitement about "developmental evolution" with its goal of explaining much of the structural diversity of organisms on the basis of (presumably physical) rules of variation at the genomic and other levels of organiza-

[35] For a history of these developments, see Keller (2002a).

[36] They have also impressed some philosophers. For instance, Fodor and Piattelli-Palmarini (2010) base part of their argument against natural selection on the basis of the existence of such laws of form.

tion and with natural selection playing much more mitigated role than in the received view of evolutionary theory.[37] What is most salient (in this context) about the project of developmental evolution is that laws of form, under the guise of laws of variation (at the genomic and, possibly, higher levels of organization) are supposed to be more important in explaining organismic (spatial) structure and variation than natural selection—but further analysis of this project is beyond the scope of this paper.

5. *A Skeptical Response*

It should not go unnoticed that all three organizational examples from Section 4 share a common feature: to varying extents, they express skepticism about the sufficiency of mechanistic explanation in biology, what I have elsewhere defended and called strong reduction (Sarkar 1998; see, also, Weber [2005]).[38] This is the idea that the behaviors of wholes, no matter how novel and unexpected they may appear to be, can be explained from the behavior of their (constituent) spatial parts (obviously including the interactions of these parts). Skepticism about this kind of reductionism, as the research programs discussed in Section 4 show, has a long pedigree in the history of biology. As emphasized several times earlier, those who deny this kind of reductionism, but wish to remain within the confines of modern (post 17th century) science, do not presume that there are processes occurring in biological (or, in general, higher structural level) systems that are not occurring in physical (or lower structural level) systems. Rather, it is a claim about explanatory adequacy or, rather, inadequacy. All research programs discussed in Section 4 share this feature.

Now, as I have contended for several decades[39], for all the fervor that it often generates, anti-reductionism (and the various associated forms of emergence) are yet to produce viable research programs with tangible content, for instance (but not limited to), predictions at variance with those made by the mundane reductionism that seems to guide almost all experimental research in biology (Weber 2005). In fact, perhaps the only positive contribution of anti-reductionism to biology, but this is an issue not contended by almost all reductionists,[40] is that reductionism provides no epistemic (or ontic, if one so chooses) warrant for eliminativism, that is, the view that reduced entities (ob-

[37] See Wagner (2007) for an entry into this literature.

[38] For expository ease, in what follows, I will call this view reductionism without the qualifier "strong." For other forms of reductionism, see Sarkar (1998) and the encyclopedia article by Brigandt and Love (2008).

[39] See Sarkar (1989, 1998, 2008) and Wimsatt and Sarkar (2006). See, also, Weber (2005).

[40] Almost all, but not all—see Churchland (1986) for a defense of eliminativism about folk psychology with respect to neuroscience. Nagel (1949, 1961) rejected eliminativism and most reductionists have (wisely) followed his lead.

jects or processes/ relations/ structures) should be replaced in scientific discourse by those used to effect the relevant reductions. The claims of this paragraph can be bolstered with plentiful and diverse cases, especially since the advent of molecular biology in the 1950s.[41] Suffice it here to mention two canonical examples relevant to teleological holism and already mentioned in Section 4.1: the allostery model which mechanistically (reductionistically) explained the co-operative behavior of macromolecules (including the Bohr effect for hemoglobin), and the operon model which so explained feedback regulation of gene expression in prokaryotes.[42] As indicated in Section 4.1, these examples are important because feedback regulation and co-operative phenomena were considered to constitute definitive exemplars of challenges to reductionism from within the anti-reductionist repertoire. Absorbing them within the reductionist agenda does much to deflate the prospects for cogent anti-reductionism.

These observations are pertinent because they help generate a strong presumption that all organizational examples of Section 4 may represent no more than flights of fancy, rich in mystical speculation about the nature and direction of biology, particularly of a future biology which remains indiscernible today, but are nevertheless devoid of empirical content. That is to say, there is no empirical basis for postulating the structures required by the teleological holists, the structural emergentists, or the developmental form theorists. It seems odd to speculate on the persistence or resilience of structures which have no empirical basis today—and worse than odd to draw strong ontological conclusions on those grounds.

Nevertheless, excessive skepticism or criticism of incipient scientific programs is also often misplaced. Take genetics. Returning to a case introduced in Section 2, if Pearson's typically highly cogent biometrical criticisms of the new Mendelism around 1900–1905 had derailed the program of Mendelism initiated by Bateson and, slightly later, Punnett, long before the advent of successful model-building by Haldane, Fisher, and Wright,[43] theoretical population genetics may well have not emerged as early as it did or, perhaps, never in the form in which it is now known and provides the basis for evolutionary theory.[44] It can, therefore, be argued that all three programs—teleological holism, structural emergentism, and developmental form theory—should be treated with tolerance, at least for the time being. Here, tolerance is supposed to mean that such research programs should not be dismissed out of hand, either epistemically (in terms of serious consider-

[41] See Sarkar (1998) and Weber (2005).

[42] For more details on these examples, see Sarkar (1998).

[43] For historical and philosophical details, see Provine (1971), and Sarkar (2004, 2007).

[44] Obviously, the second disjunct expresses some skepticism about a realist interpretation about even a body of science as empirically well-established as theoretical population genetics. This skepticism is intentional.

ation and active debate) or institutionally (in terms of funding, *etc.*), in the way, say, traditional vitalism or Intelligent Design or other forms of creationism should be so dismissed.

However, in the case of teleological holism, the time for such tolerance may well have long expired. As noted in Section 4.1, this set of claims emerged in their modern form as far back as the late eighteenth century, flourished for a while in the nineteenth century, was given new life by the physiology of the late nineteenth and early twentieth century, and reinvigorated again in the cybernetic era—all this while producing no tangible alternative to the expanding research program of resolute mechanists. The time has come to take stock of these repeated failures rather than wait for promissory notes to be delivered.[45]

Similar pessimism seems also warranted for the search for developmental laws of form. D'Arcy Thompson's *On Growth and Form* continues to provide inspiration to those who seek laws of form, and the aesthetic appeal of the book is denied by few—nevertheless it takes some faith to claim that Thompson's project any longer continues to be a useful resource for biologically-relevant inquiry (and, indeed, probably most historians of biology would now judge that it never did). Let me add that I do have that faith but my position is that of a small minority within developmental biology. To the very limited extent that models in the tradition of Turing (1952) have been successful towards the explanation of biological form, they have done so (as noted in Section 4.3) purely mechanistically, by relying on the physical (and chemical) properties of individual parts rather than mainly on structure independent of constituent details. Recent developments suggest the irrelevance of Turing-type models in contexts where they once appeared most promising, for instance, in explaining segmentation patterns in insects.[46] There appears at present to be only one prospect that may warrant tempering this pessimism—if the program of developmental evolution succeeds, and does so by explicitly going beyond standard mechanistic (reductionist) models (as, for instance, Laubichler and Wagner [2001] promise), laws of form may well enjoy a new lease of life.

This leaves the case of structural emergentism. As briefly indicated in Section 4.2, but worth special emphasis here (where a more philosophically critical appraisal of this position is being attempted), the issue of emergence is a red herring. In the present context it is not particularly interesting whether there is any interesting sense in which a feature of a system is relevantly different from those of its constituent interacting parts to be deemed emergent. What is at stake is whether in accounting for the feature, what bears the explanatory weight is the structure of the system (as modeled) compared to the identity of

[45] It is beyond the scope of this paper to assess whether a more positive—or, at least, a less negative—assessment is warranted with respect to the relationship between the mental and the biological. There is a vast philosophical literature to this topic which, fortunately, is not relevant to the topic of this paper.

[46] See Akam (1989).

the individual parts (objects). A full explication of "explanatory weight" is beyond the scope of this paper. Suffice it here to reduce it to the question whether the explanation can be extended to a large variety of other systems that have the same structure but differ in the constituent objects: the greater the differences between the sets of objects, the greater the extent to which the structure, rather than the objects, bears the explanatory weight.

The molecular explanation of dominance (which was alluded to in Section 4.2) may be one exemplar of this possibility. However, it may well be an isolated case given that no other such case seems to have been offered in the philosophical literature since Sarkar (1998) analyzed the case of dominance. Moroever, it is hard to be generally optimistic about the prospects of GRN models or any of the other kinds of network models that dominate the bulk of theoretical biology today. However, in the case of GRN models, it is too early to be sure of their eventual fate but this should surely be regarded as a situation in which excessive skepticism about an incipient research program is unwarranted. Nevertheless, all that there is at present is a promissory note.

6. *Final Remarks*

Where does this leave us? I wish to make five observations:

1. Structural realists have a wealth of evidence on their side drawn from the history of science in support of the claim that theoretical structures (for instance, relations between putative objects) are far more resilient than theoretical objects across radical theory change. This assessment is not limited to the biological contexts with which this paper is concerned. A large array of studies by (both epistemic and ontic) structural realists provide support for it from the physical sciences (Ladyman and Ross 2007).

2. Though laws (a particular type of structure) do enjoy this kind of preferential resilience compared to theoretical objects, at least in biological contexts they appear to do so only to the extent that they are phenomenological (Section 3). Moreover, even the most resilient phenomenological laws in biology do not show the degree of resilience that would warrant confidence in claims of realism about them. Section 3 showed this to be very likely in the case of the Law of Ancestral Heredity. The status of the Price equation is unlikely to be different though the verdict is still out—and will not be settled in the foreseeable future.

3. What are more likely to have the required resilience—that is, resilience not reducible to being phenomenological—are constitutive frameworks in which a variety of laws can be formulated. Recall the discussion of the Price equation in Section 3: to the extent that it seems to exhibit a high degree of resilience, it is

due to its being more akin to a constitutive framework rather than an individual law.

4. However, there is no reason to suppose that entire frameworks (including those that are at the highest level of generality as explicated in Point 3 above) may never be entirely replaced. What is troubling is that neither French (2011, 2012) nor any other structural realist seems to offer arguments to the contrary.

5. Section 4 noted that directed multigraphs may provide an appropriate constitutive framework for much of theoretical biology. Now, directed multigraphs are mathematical structures at such a high level of formal abstraction that it is important to show that the claim being made here about them is not entirely vacuous (similar, for instance, to a claim that real and complex fields [in the algebraic sense] provide a constitutive framework for physics). There are at least two restrictions that the choice of directed graphs immediately imposes: (i) Somewhat trivially, the relevant structures must exhibit some asymmetry between units (vertices) in their interactions which is represented by the directions of connecting edges. (ii) Far more importantly, directed graphs are discrete mathematical structures. Both restrictions carry over to directed multigraphs. Thus, adopting a framework of directed multigraphs assumes that biological models must be so constructed that the putative objects and relationships between them can be individuated into district sets. This excludes for instance, modeling organismic development using what used to be called morphogenetic fields, or in the way envisioned by Turing and those who followed his tradition. This means that claiming that the appropriate structures are directed multigraphs is a claim with non-trivial empirical consequences. It remains an open question whether it is correct and, if so, what other restrictions can be imposed on that structure while retaining its aim of representing as much of biological phenomena as possible.

These observations may not be much in the way of a conclusion. So, I will finally end by claiming something more definite. To the extent that this paper has defended anything at all, it has defended the importance of theoretical structures as opposed to theoretical objects. This amounts to an endorsement of *structuralism* as, for instance, explicated long ago in a mathematical context by the Bourbaki group. But it does not take any position on realism. A quote from Stein (1989: 57) is particularly relevant: "[O]ur science comes closest to comprehending the 'real', not in its account of 'substances' and their kinds, but in its account of the 'Forms' which phenomena 'imitate' (for 'Forms' read 'theoretical structures', for 'imitate', 'are represented by')." Ladyman and Ross (2007) take Stein to be sympathetic to structural realism. Arguing

against the structural realists, Stanford (2006) takes Stein (1989) to be defending a sophisticated instrumentalism. Neither of these interpretations appears to be fully accurate though Stanford's come closer (notice Stein's careful qualification "comes closest to" before any reference to the "real").

But the deeper point that Stein is making is one I would endorse and extend. Structural characterizations provide resilience against radical theory change. In particular, empirically successful phenomenological laws, interpreted as structures, not only often survive such changes but constrain the form of revised theories by being part of the data that must be accommodated. In general, these structures not only permit persistently corrected predictions (and the use of this ability for technological and other purposes) but, even as they change, they provide better representations of the world in the sense that they are resources for further enquiry that enable the extension of individual sciences and, often enough, their iterative unification. Now, what does the claim "these structures are real" or "these structures are all that can be known" add? Almost certainly, something psychological, especially for those whom James would call tender-minded (as opposed to those who are tough-minded empiricists). But it does not add anything of philosophical significance. Like Stein (1989: 65), we should maintain: realism, "yes"; but instrumentalism, "yes, also"; no only to anti-realism—with anti-realism including not only constructive empiricism but, especially, social constructivism and the other various fashionable forms of relativism that have unfortunately come to dominate much of the history of science in recent decades.

References

Akam, M. 1989. "Making Stripes Inelegantly." *Nature* 341: 282–283.

Bedau, M. A. and Humphreys, P. (Eds.). 2008. *Emergence: Contemporary Readings in Philosophy and Science*. Cambridge: MIT Press.

Bernard, C. 1865. *Introduction à l'Étude de la Médecine Expérimentale*. Paris: J.B. Baillière et fils.

Brigandt, I. and Love, A. C. 2008. "Reductionism in Biology." In Zalta, E. N. Ed. *Stanford Encyclopedia of Philosophy*. Fall 2008 Ed. http://plato.stanford.edu/archives/fall2008/entries/reduction-biology/.

Britten, R. J. and Davidson, E. H. 1969. "Gene Regulation for Higher Cells: A Theory." *Science* 165: 349–357.

Buss, L. W. 1987. *The Evolution of Individuality*. Princeton: Princeton University Press.

Churchland, P. S. 1986. *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge: MIT Press.

Davidson, E. H. 2006. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. Amsterdam: Elsevier.

Doncaster, L. 1010. *Heredity in the Light of Recent Research*. Cambridge: Cambridge University Press.

Dupré, J. 1996. "Promiscuous Realism: Reply to Wilson." *British Journal for the Philosophy of Science* 47: 441–44.

Dupré, J. and O'Malley, M. 2007. "Metagenomics and Biological Ontology." *Studies in History and Philosophy of the Biological and Biomedical Sciences* 28: 834–846.

Dupré, J. and O'Malley, M. 2009. "Varieties of Living Things: Life at the Intersection of Lineage and Metabolism." *Philosophy and Theory in Biology* 1: 1–25.

Edwards, A. W. F. 1994. "The Fundamental Theorem of Natural Selection." *Biological Reviews* 69: 443–474.

Ermentrout, G. B. and Edelstein-Keshet, L. 1993. "Cellular Automata Approaches to Biological Modeling." *Journal of Theoretical Biology* 160: 97–133.

Falk, R. and Sarkar, S. 1992. "Harmony from Discord," *Biology and Philosophy* 7: 463–472.

Fisher, R. A. 1918. "The Correlation between Relatives on the Supposition of Mendelian Inheritance." *Transactions of the Royal Society of Edinburgh* 52: 399–433.

Fisher, R. A. 1930. *Genetical Theory of Natural Selection*. Oxford: Clarendon Press.

Fodor, J. and Piattelli-Palmarini, M. 2010. *What Darwin Got Wrong*. New York: Farrar, Straus, & Giroux.

Fortuna, M. A. 2007. "Spatial Networks in Ecology." Departamento de Biología Vegetal y Ecología, Universidad de Sevilla.

Frank, S. A. "The Price Equation, Fisher's Fundamental Theorem, Kin Selection, and Causal Analysis." *Evolution* 51: 1712–1729.

French, S. 2011. "Shifting to Structures in Physics and Biology: A Prophylactic for Promiscuous Realism." *Studies in History and Philosophy of Biological and Biomedical Sciences* 42: 164–173.

French, S. 2012. "The Resilience of Laws and the Ephemerality of Objects: Can a Form of Structuralism Be Extended to Biology? In Dieks, D., Gonzalez, W. J., Hartmann, S., and Stöltzner, M. (eds.). *Probabilities, Laws, and Structures*. Berlin: Springer: 187–199.

Frigg, R. and Hartmann, S. 2006. "Scientific Models." In Sarkar, S. and Pfeifer, J. (eds.). *The Philosophy of Science: An Encyclopedia*. Vol. 2. New York: Routledge: 740–749.

Frogatt, P. and Nevin, N. C. 1971. "The 'Law of Ancestral Heredity' and the Mendelian-Ancestrian Controversy in England, 1889–1906." *Journal of Medical Genetics* 8: 1–36.

Galton, F. 1865. "Hereditary Talent and Character." *Macmillan's Magazine* 12: 157–166.

Haldane, J. S. 1906. "Life as Mechanism." *Guy's Hospital Reports* 60: 89–123.

Haldane, J. S. 1914. *Mechanism, Life, and Personality: An Examination of the Mechanistic Theory of Life and Mind*. London: John Murray.

Heims, S. J. 1991. *Constructing a Social Science for Postwar America: The Cybernetics Group, 1946–1953*. Cambridge: MIT Press.

Holton, G. 1970. "The Roots of Complementarity." *Daedalus* 99: 1015–1055.

Horgan, J. 1995. "From Complexity to Perplexity: Can Science Achieve a Unified Theory of Complex Systems?" *Scientific American* 272 (6): 104–109.

Jerne, N. K. 1974. "Towards a Network Theory of the Immune System." *Annales d'Immunologie* 125C: 373–389.

Johannsen, W. L. 1905. *Arveligheds Laerens Elementer*. Copenhagen: Nordisk Forlag.

Keller, E. F. 2002a. *Making Sense of Life: Explaining Biological Development with Models, Metaphors, and Machines*. Cambridge: Harvard University Press.

Keller, E. F. 2002b. *The Century of the Gene*. Cambridge: Harvard University Press.

Koonin, E. V. 2012. *The Logic of Chance: The Nature and Origin of Biological Evolution*. Upper Saddle River: Pearson Education.

Ladyman, J. and Ross, D. 2007. *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Laubichler, M. and Wagner, G. P. 2001. "How Molecular is Molecular Developmental Biology? A Reply to Alex Rosenberg's Reductionism Redux: Computing the Embryo." *Biology and Philosophy* 16: 53–68.

Lenoir, T. 1989. *The Strategy of Life: Teleology and Mechanics in Nineteenth-Century German Biology*. Drodrecht: Reidel.

Lewin, B. 1974. *Gene Expression. Volume 2. Eucaryotic Chromosomes*. New York: Wiley.

Lock, R. H. and Doncaster, L. 1920. *Recent Progress in the Study of Variation, Heredity, and Evolution*. 6th. Ed. London: John Murray.

Lynch, M. 2007. *The Origins of Genome Architecture*. Sunderland: Sinauer.

Monod, J. 1971. *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*. New York: Vintage.

Moran, P. A. P. and Smith, C. A. B. 1966. "Commentary on R. A. Fisher's 'The Correlation between Relatives on the Supposition of Mendelian Inheritance." *Eugenics Laboratory Memoirs* 41: 1–62.

Nagel, E. 1949. "The Meaning of Reduction in the Natural Sciences." In Stauffer, R. C. (ed.). *Science and Civilization*. Madison: University of Wisconsin Press: 99–135.

Nagel, E. 1951. "Mechanistic Explanation and Organismic Biology." *Philosophy and Phenomenological Research* 11: 327–338.

Nagel, E. 1952. "Wholes, Sums, and Organic Unities." *Philosophical Studies* 3: 17–32.

Nagel, E. 1961. *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace, and World.

Newman, S. E. 2019. "Cell Differentiation: What Have We Learned in 50 Years?" https://arxiv.org/abs/1907.09551.

Olby, R. C. 1966. *Origins of Mendelism*. New York: Schocken Books.

Olby, R. C. 1987. "William Bateson's Introduction of Mendelism to England: A Reassessment." *British Journal of the History of Science* 20: 399–420.

Pascual, M. and Dunne, J. E. 2006. *Ecological Networks: Linking Structure to Dynamics in Food Webs*. New York: Oxford University Press.

Pearson, K. 1900. *The Grammar of Science*. 2nd. ed. London: A. and C. Black.

Pearson, K. 1904a. "Mathematical Contributions to the Theory of Evolution. XII. On a Generalized Theory of Alternative Inheritance, with Special Reference to Mendel's Laws." *Philosophical Transactions of the Royal Society (London) A* 203: 53–86.

Pearson, K. 1904b. "A Mendelian's View of the Law of Ancestral Inheritance." *Biometrika* 3: 109–112.

Perini, L. 2011. "Sequence Matters: Genomic Research and the Gene Concept." *Philosophy of Science* 78: 752–762.

Price, G. R. 1972. "Fisher's 'Fundamental Theorem' Made Clear." *Annals of Human Genetics* 36: 129–140.

Provine, W. B. 1971. *The Origins of Theoretical Population Genetics*. Chicago: University of Chicago Press.

Raff, R. A. 1996. *The Shape of Life: Genes, Development, and the Evolution of Animal Form*. Chicago: University of Chicago Press.

Sarkar, S. 1989. "Reductionism and Molecular Biology: A Reappraisal." PhD Dissertation. Department of Philosophy. University of Chicago.

Sarkar, S. 1992a. "Haldane as Biochemist: the Cambridge Decade, 1923–1932." In Sarkar, S. (ed.). *The Founders of Evolutionary Genetics*. Dordrecht: Kluwer: 53–81.

Sarkar, S. 1992b. "Science, Philosophy, and Politics in the Work of J. B. S. Haldane, 1922–1937." *Biology and Philosophy* 7: 385–409

Sarkar, S. 1996. "Biological Information: A Skeptical Look at Some Central Dogmas of Molecular Biology." In Sarkar, S. (ed.). *The Philosophy and History of Molecular Biology*: *New Perspectives*. Dordrecht: Kluwer: 187–231.

Sarkar, S. 1998. *Genetics and Reductionism*. New York: Cambridge University Press.

Sarkar, S. 1999. "From the *Reaktionsnorm* to the Adaptive Norm: The Norm of Reaction, 1909–1960." *Biology and Philosophy* 14: 235–252.

Sarkar, S. 2004. "Evolutionary Theory in the 1920s: The Nature of the Synthesis." *Philosophy of Science* 71: 1215–1226.

Sarkar, S. 2007. "Haldane and the Emergence of Modern Evolutionary Theory." In Matthen, M. and Stephens, C. (eds.). *Handbook of the Philosophy of Science. Volume 3: Philosophy of Biology*. New York: Elsevier: 49–86.

Sarkar, S. 2008. "Reduction." In Psillos, S. and Curd, M. (eds.). *The Routledge Companion to the Philosophy of Science*. London: Routledge: 425–434.

Schaffner, K. S. 1974. "Logic of Discovery and Justification in Regulatory Genetics." *Studies in History and Philosophy of Science* 4: 349–385.

Smuts, J. C. 1926. *Holism and Evolution*. New York: MacMillan.

Stanford, K. 2006. *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. New York: Oxford University Press.

Stein, H. 1989. "Yes, but . . . Some Skeptical Remarks on Realism and Antirealism." *Dialectica* 43: 47–65.

Stern, C. 1965. "Mendel and Human Genetics." *Proceedings of the American Philosophical Society* 109: 216–226.

Thompson, W. D. 1917. *On Growth and Form*. Cambridge: Cambridge University Press.

Tigerstedt, R. 2012. "Christian Bohr: Ein Nachruf." *Skandinavisches Archiv für Physiologie* 25 (2): v–xviii.

Turing, A. M. 1952. "The Chemical Basis of Morphogenesis." *Philosophical Transactions of the Royal Society (London)* 237: 37–72.

van Fraassen, B. C. 1980. *The Scientific Image*. Oxford: Oxford University Press.

Wagner, G. P. 2007. "The Current State and the Future of Developmental Evolution." In Maienschein, J. and Laubichler, M. Eds. *The History of Evolutionary Developmental Biology*. Cambridge, MA: MIT Press: 525–545.

Weber, M. 2005. *Philosophy of Experimental Biology*. New York: Cambridge University Press.

Weinberg, W. 1909a. "Über Vererbungsgesetze beim Menschen. 1. Allgemeiner Teil." *Zeitschrift für Induktive Abstammungs- und Vererbungslehre* 1: 377–392, 440–460.

Weinberg, W. 1909b. "Über Vererbungsgesetze beim Menschen. 2. Spezieller Teil." *Zeitschrift für Induktive Abstammungs- und Vererbungslehre* 2: 276–330.

Weinberg, W. 1910. "Weitere Beiträge zur Theorie der Vererbung." *Archiv für Rassen- und Gesselschafts-Biologie* 7: 35–49, 169–173.

Wimsatt, W. C. and Sarkar, S. 2006. "Reductionism." In Sarkar, S. and Pfeifer, J. (eds.). *The Philosophy of Science: An Encyclopedia*. Volume 2. New York: Routledge: 696–703.

Worrall, J. 1989. "Structural Realism: The Best of Both Worlds?" *Dialectica* 43: 99–124.

Wright, S. 1930. [Review of Fisher, R. A. 1930. *Genetical Theory of Natural Selection*. Oxford: Clarendon Press.] *Journal of Heredity* 21: 349–356.

Yule, G. U. 1902. "Mendel's Laws and Their Probable Relations to Intraracial Heredity." *New Phytologist* 1: 193–207, 222–238.

Croatian Journal of Philosophy
Vol. XX, No. 58, 2020

Does Sherlock Holmes Exist?

RICHARD VALLÉE
Université de Moncton, Shippagan, Canada

Fictional names have specific, cognitively relevant features, putting them in a category apart from the category of ordinary names. I argue that we should focus on the name or name form itself and refrain from looking for an assignment procedure and an assigned referent. I also argue that we should reject the idea that sentences containing fictional names express singular propositions. These suggestions have important consequences for the intuition that 'Sherlock Holmes exists' is either true or false, and they put our intuitions concerning fictional names into perspective. If Millianism is the view that names only have a referent only as their semantic value, then my proposal on fictional names is not Millian in nature.

Keywords: Fictional names, existence, pluri-propositionalism, cognitive significance.

1. *Fictional Names and Existence*

It is widely assumed that 'Sherlock Holmes' is a fictional name, that is, a name introduced in fiction, and a name with no referent in the real world (see Kripke 1980; Kripke 2011; Kripke 2013).[1] If a fictional name such as 'Sherlock Holmes' originates in fiction and has no three-dimensional referent located in space and time then, on the orthodox analysis of names, the name lacks a referent. *Prima facie,* it has no semantic value. As a consequence 'sentences containing it say nothing' (Braun 1993: 449). However, intuitively sentences such as

(1) Sherlock Holmes exists

are true or false, and convey information we can agree or disagree about; e.g., you and I may disagree about whether or not (1) is true.

[1] There is a distinction to be made between fictional names, like 'Sherlock Holmes', and names of imaginary friend, or fantastic creature ('Nessie'), the speaker wrongly believes to exist and to be the referent of that name. My paper concerns fictional names only.

You can even argue that

(2) Sherlock Holmes does not exist.

is true. However, just like (1), (2) says nothing and expresses no proposition. Thus, it has no truth conditions and it is neither true nor false. I am interested in singular existential and negative singular existential sentences containing fictional names, like (1) and (2). The question, then, is in determining how such sentences can be truth-apt despite the fact that, intuitively, 'Sherlock Holmes' lacks a real world referent. Furthermore, how can someone believe that (1), or (2), is either true or false if 'Sherlock Holmes' has no referent in the real world? Why do we disagree if 'Sherlock Holmes' has no referent? What is our disagreement about? Problems concerning fictional names are, arguably, semantic in nature, concerning key semantics notions such as reference and truth.[2] People have a strong inclination to model fictional names on ordinary names—i.e., to think about 'Sherlock Holmes' as analogous to ordinary proper names such as 'Barack Obama', and are subsequently tempted to assign referents to fictional names as well. Following the now orthodox view on names, an utterance of (1), or (2), expresses a singular proposition just as an utterance of 'Obama exists' expresses the singular proposition <OBAMA, exists>, which contains Obama himself as a constituent. The problem is that there is purportedly no such thing as Sherlock Holmes to be introduced into a proposition (see Braun 2005; Adams 2011). Many philosophers have tried to account for fictional names (see Kripke 2013; Currie 1990; Walton 1990; Thomasson 1990; Braun 2005; Adams 2011; Kroon 2014). In section 3, I will sketch and criticize two main, paradigmatic perspectives on such names and fictional sentence—a 'pretense' perspective (Kripke, Walton) and an empty proposition perspective (Braun).[3]

Ordinary and fictional names differ in many important ways. In contrast with ordinary names, fictional names are not assigned to individuals by ordinary speakers the way that 'Barack Obama' is. Moreover, fictional names should not be characterized simply by the fact that they lack a three-dimensional referent in the real world, which can be a constituent of a singular proposition. To give them a purely negative characteristic only lays ground for an oversimplified picture of such names. Fictional names have specific, cognitively relevant features, putting them in a category apart from the category of ordinary names. In section 2, I introduce ordinary names and fictional names, and two problems raised by fictional names. Section 3 briefly discusses,

[2] There is an important literature on fictional names and existence, e.g. Kripke (2011; 2013); Walton (1990; 2000); Braun (2005); Everett (2003; 2007). Everett and Hofweber's *Empty Names, Fiction and the Puzzle of Non-Existence* provides a good collection of articles on the problem.

[3] There are too many different views on fictional names and existence to deal with in a short paper. Moreover, my paper is not intended as a criticism of these views, but as a new, modest contribution to what remains a puzzling issue.

first Kripke's suggestion invoking pretense. Calling it pretense suggest that it is not literal. If I am right, singular existential sentences are literal, and the idea of pretense should be dispensed with. I discuss, second, Braun's view. The point here is to reject views according to which fictional names make room for an object in a proposition, while leaving this room not filled. Section 4 introduces the framework I use, pluri-propositionalism. Section 5 sketches my perspective on what fictional names are. Section 6 applies pluri-propositionalism to fictional names in existence sentences, and offers solutions to the problems presented in section 3. Section 7 concludes the paper. I argue that we should focus on the name or name form—that is, its written form—itself and refrain from looking for an assignment procedure and an assigned referent. I also argue that we should reject the idea that sentences containing fictional names express singular propositions. These suggestions will have consequences for the intuition that (1) is either true or false, and put our intuitions concerning fictional names into perspective. If Millianism is the view that names have only a referent as their semantic value, then my proposal on fictional names is not Millian in nature. I argue that by abandoning the analogy perspective certain problems raised by existence sentences such as (1) and (2) can be addressed in a novel way.

From the perspective of someone reading a fictional story containing fictional names, fictional names behave like regular names. This is part of what explains why reading fiction is an enjoyable activity. Fiction readers use their imagination, as they read the fictional story, to create a picture of what the individual designated by such a name would be like. I read Hammett's books, and I have a picture of Sam Spade; you read the same book and you (most) probably have a different picture of Sam Spade. However, outside of the enjoyment that it provides while reading fiction, it is not really appropriate in semantics to see fictional names as ordinary names. Discussions about existence sentences such as (1) usually take fictional names to be just that, names, thereby disregarding the aspects that make them fictional names—aspects which distinguish them from ordinary names such as 'Barack Obama', for instance. Standard analyses place too much emphasis on 'exists' rather than on the fictional name itself in (1). Locutions such as 'exists in fiction' are frequently invoked to avoid an array of problems that are raised by 'exists' where fictional names are involved. My approach does just the opposite: it emphasizes fictional names.[4] The category of a fictional name deserves special attention, and it can be characterized in an epistemically fruitful way, echoing speaker's intuitions. Once fictional names are considered, 'exists'—as well as 'exists in fiction'—takes a back seat and is innocuous.

[4] With, as a result, evading complex issues concerning existence examined by Predelli (2002).

2. *Ordinary and Fictional Names*

According to the Theory of Direct Reference (Kripke 1980), the sole semantic function of a name, e.g., 'John Smith', is to refer to an individual, e.g., John Smith, and a name's sole semantic value is the individual it is assigned to. The general principle underlying this view is the assignment of a value to a variable: the latter being the sequence of sounds or the sequence of letters playing the role of a name, while the relevant value is the bearer of such a name, e.g., John Smith. On this picture, the sentence 'John Smith is a detective' expresses a singular proposition which contains the individual the name is assigned to (i.e., John Smith) and the property of being a detective as constituents: < JOHN SMITH, being detective>. Clearly, many people bear the name 'John Smith'. Although it may seem as though a single name gets different individuals assigned to it as its various values, it is arguable that we, in fact, have different names. According to this theory, multiple instances of 'John Smith' are to be counted as homonymous expressions with each name being individuated by its semantic value (i.e., the individual). A sentence containing the name 'John Smith' expresses a singular proposition containing the specific individual that the name is assigned to. This singular proposition provides the truth conditions of the sentence. If a phonetically identical name is assigned to a different person, then there are two different names, which contribute two different referents to the propositions expressed by sentences containing them. These sentences express two different singular propositions and therefore have distinct truth conditions.

Philosophers of language usually address the problem of fictional names directly, and usually treat them as names just like any other names. If fictional names are similar to ordinary names but lack a referent, then analyzing them like ordinary names via the direct reference paradigm leads to the result that the proposition that the sentence determines is, at best, the incomplete proposition: < , exists>. An affirmation of the existence of Holmes does not, then, make much sense. Nevertheless, the question of whether or not Sherlock Holmes exists seems to persist. Moreover, if a fictional name has no semantic value, then it cannot be individuated by it. It then becomes difficult to draw a distinction between the name of Doyle's famous 19th century detective and the name of his also famous 21st century counterpart in a television series. Which Sherlock Holmes is the existential question about? Is it about the 19th century detective, or about its 21st century counterpart? Or is it about another Sherlock Holmes? Unless further details are added, these questions remain difficult to answer, if they can be answered at all. Fictional names, in contrast with names such as 'Vulcan', do not seem to lack a referent because a mistake happened when assigning the name. Fictional names are, arguably, not designed to designate an object located in space and time—that is why the name is fictional to begin with (Kripke 2013). In this respect, one cannot re-

ally say that a fictional name 'fails to refer' since there is no object it is supposed to refer to in the first place. Fictional names are not mere empty names. I will address fictional names in an indirect way in section 5.

Many competent speakers are inclined to judge both (1) and (2) as true. Herein lies a puzzle: assuming that we are talking about the same Sherlock Holmes, *prima facie* our intuitions to judge both (1) and (2) to be true leads to a contradiction, since Sherlock Holmes cannot exist and not exist at the same time. However, perhaps the puzzle is misguided, since 'Sherlock Holmes' does not designate anything. Let's call this the contradiction problem. Some may argue that there is a sense in which Sherlock Holmes does exist and another sense in which he does not. But this is just a description of the problem.

Plausibly, Donald Trump believes that Sherlock Holmes does not exist. One can then ask what he believes exactly if 'Sherlock Holmes' is a fictional name. One can also ask which Sherlock Holmes is his belief about? Is Trump's belief about the Doyle character in his 19th century novel or about the character in the television series? Or about both, assuming that they are one and the same? In any case, his belief is not about a three-dimensional, real individual. Presumably Obama also believes that Sherlock Holmes does not exist. Which Sherlock Holmes is his belief about? Do Trump and Obama share a belief in common? Let's call this the belief attribution problem. Intuitions about the truth conditions of existential statements about fictional characters are complex (see Braun 1993; Thomasson 2003; Predelli 2002). Braun (2005) is right in noting that speakers have a cognitive relationship to fictional names 'that [is] importantly similar to the cognitive relations they bear to referring names' (Braun 2005: 600). However, Braun also suggests that they are not entirely analogous. Neither (1) nor (2) are made true by facts, since *prima facie* facts drop out of the picture as far as sentences containing fictional names are concerned. Existential statements, such as (1) and (2), need to be more aptly analyzed.

3. *Using the Ordinary Name Paradigm*

There is an important literature on fictional names invoking pretense (see for example Kripke (2013); Walton (1990; 2000); Kroon (2014)). Call it the pretense family type of theory. In *Reference and Existence* (2013), Kripke writes that 'the type of names which occurs in fictional discourse are pretended names', and that 'the propositions in which they occur are pretended proposition rather than real propositions' (Kripke 2013: 29). The speaker does not, then, literally refer to an object or express a proposition. The speaker only pretends to use the name and to express a proposition. Such a view does not entail that pretend names refer to fictional objects.[5] Suppose now that fictional names are

[5] Let's suppose that fictional names do refer to fictional objects, a view which

pretended names, and that sentences containing such names do not determine propositions, but just pretended propositions. We are owed details on what a pretended name is, and on what pretended propositions are. An approach to fictional names preserving the intuitions that such names are actual names, not pretended names, and that such sentences express propositions, not pretended propositions, would have much in its favour. Walton (1990) also uses the notions of pretense, as well as the notion of make-belief, and resists the intuition that (1) and (2) are literal, or used literally. He offers a very rich view on fictional name sentences. I will not offer detailed criticisms of Walton's picture. My view dispenses with the notion of pretense and takes sentences containing fictional names to be literal.

A different strategy, or family of strategies, is to argue that fictional names are, in fact, referring expressions, but that they refer to nothing. The way out is to argue that such names lack referent, and that a sentence like (1) expresses a gappy proposition, e.g., <___ , exists>, (Braun 2005). Braun argues forcefully for this position, and he suggests what the truth conditions are for both (1) and its negation, here (2). He would contend that the gappy proposition determined by (1) is false, and that its negation is true (see Braun 2005: 599).[6] Falsity is used in an odd sense here, in that it does not consider facts. Assigning falsity to (1) seem arbitrary. On this view, a sentence such as 'Donald Trump believes that Sherlock Holmes exists' is a belief report containing a gappy proposition. If the gappy proposition is false, and if its negation is therefore true, then the sentence 'Donald Trump believes that Sherlock Holmes does not exists' is true, and it attributes to Trump a true belief. This does not seem correct.[7] Braun's picture, leave many questions unanswered. According to his view, different sentences containing different fictional names—'Sherlock Holmes exists', 'Philip Marlow exists' and 'Martin Beck exists'—determine the same gappy proposition and share the same truth conditions. Important differences are obliterated. In addition, 'Sherlock Holmes is not Philip Marlow' and 'Martin Beck is not Sherlock Holmes', as well as 'Sherlock Holmes is not Martin Beck', determine the same gappy proposition: < ___ is not ___ >. *Prima facie*, they determine different propositions and have dif-

has been defended quite strongly in the literature (see Thomasson 1990). However, it is often assumed that such fictional characters either exist or do not exist. This assumption just begs the question, and it is not a satisfactory response to puzzles such as (1) and (2). We are also owed an account of how fictional names are assigned to such fictional objects, whatever the latter are supposed to be Moreover, if the name refers to a fictional object, then the truth conditions of (1), or an utterance of (1), for instance, then remain puzzling. How can a fictional object make a sentence true, in a non pickwikian sense of 'true'? Are (1) and (2) contradictions? For a critical perspective on Thomasson'view, see Everett (2007).

[6] Adams (2011) argues for gappy propositions, but in contrast with Braun, he contends that gappy propositions, determined by a sentence like (1), are neither true nor false. If he is right, then (1) and (2) are not contradictions.

[7] I will not examine Braun's proposal in detail here.

ferent truth conditions. Braun's view seems to imply, against intuitive judgements, exactly the opposite: that they determine the same gappy proposition.

Most of the various approaches to fictional names address issues concerning their reference and their contribution to the truth conditions of sentences containing them. They are mainly designed to deal with the truth-value problem. I will not try to give details on how such approaches deal with the contradiction problem and the belief attribution problem. Other important issues can also be raised. For example, do fictional names contingently lack a referent? What are the features of fictional names that make these expressions referential terms, which nevertheless lack a referent? Such lack of referent is *prima facie* not accidental (see Kripke 1980). If 'Sherlock Holmes' is found to have a referent, then that name does not count as a fictional name. Being fictional is arguably not a contingent feature of fictional names. Finally, different uses of 'John Smith' can be individuated distinctly so long as the name is tied to different objects. Different 'Sherlock Holmes' can be found in Doyle's book, in movies and in the television series. Are they different names? If they lack referents then they cannot be individuated. To avoid the problems that these questions raise, some philosophers, such as Braun, suggest taking the modes of presentation, the names themselves, 'Sherlock Holmes', used to express that proposition as relevant for distinguishing the proposition and truth conditions associated with each sentence or utterance. Such a procedure, however, cannot distinguish some pairs of fictional names. Suppose that, impressed by Conan Doyle's books, I decide to write mystery novels, and Berlin in 2018 is a great place for a mystery. I also decide to call my German detective 'Sherlock Holmes.' Now, two tokens or utterances of the sentence (1), one about Conan Doyle's detective and the other about the Berlin cop, determine the same gappy proposition. They *prima facie* also contain the same name, i.e., 'Sherlock Holmes'. The name, by itself, is useless in distinguishing what proposition is determined by each sentence or utterance. Such a result is counterintuitive. A more fine-grained individuation of fictional names is called for, one not invoking any fictional character. Braun, for instance, addresses neither the identity problem for fictional names nor the problem of the multiplicity of similar fictional names like 'Sherlock Holmes' in Doyle's books or in a television series. The individuation of fictional names is puzzling, yet it is also relevant in addressing sentences such as (1) and (2).

In a passage I quoted earlier, Braun (2006) suggests that the cognitive relations ordinary speakers have to fictional names are not exactly the same as the relations they have to fictional names when the speakers know that a name is fictional. On this point I believe that Braun is right. Linguistically competent or informed speakers know that there is a difference between ordinary and fictional names—e.g., that there is a difference between the name 'Frank Serpico' and the name 'Sherlock

Holmes'. The first one is a real name, and the second one is a fictional name. They also know that 'Frank Serpico is a detective' and 'Sherlock Holmes is a detective' are, in some way, at odds. Consider an utterance such as 'Frank Serpico is a detective, and so is Sherlock Holmes'. Such an utterance is quite complex. The difference between these names does not simply trace back to their referential aspects. Features of fictional names make it such that they do not refer to objects. In addition, fictional names have a specific cognitive impact on speakers, an impact quite distinct from the cognitive impact of ordinary names. Before examining fictional names, let me introduce the pluri-propositionalist framework that I will be using. Such a framework provides interesting ways to address the issues raised by fictional names.

4. *The Many Truth Conditions*

In philosophy of language, the idea that each sentence or utterance determines one single proposition, or mono-propositionalism, is paradigmatic. It is found in Frege's pioneering work and in most of the subsequent views. Pluri-propositionalism offers a different perspective on sentences and utterances. Following pluri-propositionalism (see Perry 2012; Korta and Perry 2011), utterances rather sentences are in the foreground. In this respect, pluri-propositionalism focuses on linguistic communication. It also argues that many different propositions or truth conditions can be determined by a single utterance of a sentence. Linguistic expressions have linguistic meaning, that is, a rule determining the content of utterances of those expressions. Meaning fixes the semantically determined proposition, content, or truth conditions of utterances.[8] Consider for example an utterance of

(3) Meryl Streep exists.

The utterance, **u,** of (3) is individuated by the speaker, the time, say May 16, 2018, and the location of the utterance, say San Francisco. 'Meryl Streep' is an ordinary proper name. Following the Theory of Direct Reference, it has no linguistic meaning and a referent only. The name is associated with a referent by a convention. These features are echoed in an understanding of the utterance. Being linguistically competent and relying on their knowledge of language only, including their knowledge of what a proper name is, speakers know that:

> Given that (3) is an English sentence, the utterance **u** of (3) is true if and only if the individual[9] that the convention exploited by the utterance **u** allows us to designate by 'Meryl Streep' exists.

[8] For simplification, I do not make a difference between spoken and written utterances of a sentence or a name.

[9] An individual is whatever is designated by a proper name.

Meryl Streep does not have to exist to obtain such content. The speaker does not even have to know who Meryl Streep is. The content of the semantically determined truth conditions of the utterance, without considering facts, accounts for the cognitive significance of the utterance (Perry 2012). Note that the name itself is part of the cognitive significance of the utterance. Different utterances of (3) have different cognitive significance, because each contains a different utterance. Such contents can be accepted as true. Accepting such contents as true is an attitude toward the utterance or the content of the utterance. The latter itself cannot be said to be true, because facts have not been introduced. If the cognitive significance classifies as an episode of thinking, we can take the latter to be in the speaker's head. Such content contains the utterance **u** itself as a constituent and is, hence, reflexive in relation to the utterance itself. I underline the fact that the name itself, 'Meryl Streep', is mentioned in the cognitive significance of the utterance giving it, as an object, a major cognitive role. What follows 'if and only if', and precedes 'exists', captures an important aspect of the reference or designation relation. Yet, what you then understand when hearing an utterance of (3) does not depend on the referent of the name on that particular utterance. Actually, at this stage the referent plays no role at all. The name is associated with a convention tying it to a real individual, MERYL STREEP herself, that is: after taking into account facts required for fixing the designation of referring terms,

> Given that (3) is an English sentence, the utterance **u** of (3) is true if and only if MERYL STREEP exists.

'MERYL STREEP *exists'* is the designational content of the utterance, giving the conditions under which the utterance is true given the facts. The designational content of the utterance of (3) does not contain the utterance of that sentence. Neither does it contain the name. The designated individual has that feature, i.e., it exists or not, independently of whether or not there is an utterance at all, and whether or not that name has been assigned to that individual. All utterances of (3) with that specific name associated with the same convention have the same designational content and, given the facts, are true. In contrast with ordinary names, fictional names do not designate real objects, and *prima facie* an utterance of (1) does not have designational content. Therefore, they do not introduce anything to the truth conditions of sentences or utterances of which they are a part. Yet, the cognitive significance of utterances containing fictional names remains on the table.

5. *Introducing Fictional Names*

It is assumed that fictional names have no referent in the real world. It is also widely agreed that there are no ordinary speakers introducing them into discourse. One must make a clear distinction between the author of the fiction, Conan Doyle, Dostoievky, or Marcel Proust

for instance, and the fictive narrator, the latter introducing fictional names in the fiction itself. Such a distinction is part of the tools used by authors, and it is standard in the literature on fiction. Doyle, the author, created the Holmes stories; Watson, the fictive narrator, tells them and introduces, and uses as well, 'Sherlock Holmes'. Doyle never met Watson, and vice-versa. If it were discovered that a novel, previously thought to have been told by a fictive narrator—Watson—turned out to be the work of the author himself—Doyle—it would not be fictitious novel anymore. If Doyle's mysteries were, in fact, autobiographical, they would not be fictional. In the philosophy of language, the distinction between the author and the fictive narrator is not always clearly made or it is simply ignored altogether. My suggestion takes it into account.

Whereas ordinary names are simple entities lacking meaning, and are individuated by a sequence of phonemes and an object assigned to it by another three-dimensional creature, fictional names lack both meaning and a referent. There is no name assignment of a fictional name by a real person. Fictional names are certainly individuated by a sequence of sounds or letters. But that is clearly not enough, and it is insufficient to distinguish the name of the famous Doyle detective and, say, the name of a character in a different fiction. Nonetheless, readers individuate fictional names, and can see a difference between 'Sherlock Holmes', the name of the famous detective, and 'Sherlock Holmes', the name of a different character, say, a sailor in a fiction. Furthermore, if no fictive narrator introduces a fictional name in fiction, then there is no fictional name at all. If Watson does not use 'Sherlock Holmes', it is difficult to say something about Sherlock Holmes. An author using created names, without fictive narrators, is not introducing fictional names. Arguably, fictive narrators can also use the name of real people, or ordinary names, like 'Aristotle' or 'Ludwig Wittgenstein', despite the fact that these are not fictional names. I propose to individuate fictional names by taking the source of the name in fiction into account -Watson in a Doyle mystery for instance, and any book will do here. I do not want to explore the specifics of particular fictional names. My aim is more general and, up to a point, detached from literary theory.

The name 'Sherlock Holmes' is introduced by Watson, a fictive narrator, in Doyle's mysteries, and 'William de Baskerville' is introduced by Adso of Melk, also a fictive narrator, in Eco's *The name of the rose*. I want to draw attention to the fact that neither Kripke, nor Thomasson or Braun, the main contemporary theorists concerned with fictional names, gives fictive narrators a role. They all take only the actual authors into consideration. Philosophers of language ignore the semantic impact of what is a major literary element. The name of the author, or the title of the book, is sometimes used to focus on and identify the relevant fictional name. Real objects are then relied on to zoom in on a specific fictional name and, as things happen, on a character. That

does not imply that it—the title of the book for example—individuates that name. Fictional names are not introduced in the usual way that ordinary names are in language. A fictional name is never assigned a three-dimensional entity having causal relationships with other objects because, obviously, the fictive narrator, being fictive, has never met any.[10] A fictional name is also individuated in a very specific way. Let me explain.

Consider the sentence

(4) Holmes was certainly not a difficult man to live with.

'Holmes' can be an ordinary name (e.g., the name of a London taxi driver) or a fictional name. A token of (4) retains no information about its origins. Tokens of names, like those in (4), can be found in very different places, including Chinese cookies, books, and songs. An utterance is an event, which is located in space and time, and is the production of a sentence by a speaker. An utterance of (4) keeps no trace of its speaker, nor its space and time parameters. A sentence such as (4) can be uttered and unless you are in contact with the original utterance, these parameters remain unknown. As a consequence, no one can tell whether 'Holmes' is a fictional name or not. Now, if I tell you that that token came from a mystery novel, *A Study in Scarlet,* by Arthur Conan Doyle, you are informed that it is a token originating in fiction. Let's call this the source of the fictional name. You know something extralinguistic about the token of the name. The sentence is reproduced in different copies of the book, and you can confidently expect to find that sentence, or more precisely tokens of that sentence, in every copy of that book. It is definitely not like an ordinary token. Common sense suggests that this is not an autobiography and that Doyle is not the fictive narrator. Maybe you remember your literature classes and you know that Watson is the narrator. If you do not, you just call the storyteller 'the narrator'. If you know a little bit more about that piece of literature, you also know that 'Holmes' is a fictional name. This is part of your knowledge of that name. You also know that that name, in every token of that sentence, in every copy of the book, is a fictional name. Fictional names have identifying features, which trace back to their source—e.g., a Doyle mystery. That token, and the name, was not designed to be located in the 17th century, and it was not designed to come from China or Japan. Watson, a fictive narrator, wrote in London at the end of the 19th century, and you read it in *A Study in Scarlet*. Neither (1) nor (2) are specific about the name 'Sherlock Holmes' they contain. There is different fictional 'Sherlock Holmes'. I will come back to it.

The sentence (4) is a fiction sentence. The sentence is indexed to a narrator, the fictional Watson, a time (end of 19th century), and a location (i.e., London). Call this its fictional index. These are part of what

[10] One can easily imagine a fiction without fictional names. Just substitute a definite or an indefinite description to each and every fictional name, in *A Study in Scarlet* for example.

makes it is a fictional sentence. The fictional index gives individuating features of the fiction sentence.[11] The fiction sentence keeps its fictional index even in different copies of the book, in graffiti, written on pieces of paper in Chinese cookie, etc. Different tokens of that sentence share the same fictional index. I think that fictional names are complex objects individuated by a sequence of letters and the fictional index, composed of the fictive narrator, time, and location of the sentence token wherein the name is introduced in fiction: e.g., 'Sherlock Holmes' (Watson, end of 19th century, London).[12] The name coming from (4) is individuated by the fictional index of the token it is taken from. Neither the location nor the time need be very specific. If the specific index is not given (or known), we can have: (narrator of **t**, time of **t**, location of **t**).[13] **T** is a token of the fictional name found in the fiction. The index gives the source of the name. I call it the indexed token of the fictional name. Names like 'Vulcan', not coming from fiction, are not assigned a fictional index. A fictional name can be extracted from a fictional sentence with its index, and then proceeds to have a life of its own. For example, 'Sherlock Holmes', 'Raskolnikov', 'Madame de Villeparisis' and 'William of Baskerville' are all famous fictional names coming from fictions and having a life of their own. Sometimes the narrator has a name, Adso of Melk for instance, and a time as well as a place, are provided. Time and location can be fuzzy and not very specific—a year or a city—as is usually the case in fiction. Readers often do not even know who the fictional narrator is. In the same way, they do not care much about the time of the writing. They also very often ignore the location of the writing. Except when they are essential to an understanding of the story, these features are just irrelevant to most readers. Hence, the narrator of the name, the time and the location of the writing are used only sometimes. In any event, the fictional status of the name 'Sherlock Holmes' is echoed in the truth conditions of the utterance of (1). That being said, 'Watson' is also a fictional name and it needs to be indexed as well: e.g., 'Watson' <Watson, London, end of 19th century>. Being indexed is a characteristic of fictional names. Given the scope of the present paper, I will set aside the issue of indexing the name used by the fictive narrator.

A fictional name can be used outside of the context of the fictional work that it originated from, while nevertheless keeping its identifying fictional index as in an utterance of (1). All fictional name tokens have a fictional index. If a name does not, then the name refers directly

[11] In that respect, its role is very different from, and should not be confused with, the role of indexes in theories of indexicals. Fictional indexes are introduced and used to account for fictional names in Vallée 2018.

[12] I give a minimal fictional index, and leave it as an open question whether more indices should be added.

[13] Different sequences of phonemes should be considered because in different languages (Russian, Japanese, French, and so on) names are written and pronounced in different ways. I put aside this issue for now.

to its designata by default. Knowing that a name is fictional means knowing that it has a fictional index, and vice-versa. In day-to-day casual conversation about Holmes the index is mostly left unspecified and remains in the background. Consequentially, casual conversations involving fictional names are occasionally unclear, since they involve utterances which have no well determined truth conditions. As a result, sometimes it is necessary to pause the conversation in order to clarify which 'Sherlock Holmes' is relevant. Different tokens of a fictional name sentence such as 'Sherlock Holmes is a detective', whether they are found on a school wall, a London bus or a piece of paper in a Chinese cookie, do not indicate whether the name used is a fictional name, the name of a London detective, or the name of an American cowboy. Regardless, the name is indexed, and this is what makes it the specific fictional name as found in Doyle's mystery, for example.

6. *On Fictional Names and Pluri-Propositionalism*

Let us return to the questions concerning fictional names and existence. Mono-propositionalism is a framework assumed by all currently proposed theories of fictional names. Pluri-propositionalism provides a new perspective through which we can examine them. Consider a sentence such as (1). Without specifying whether or not it is a fictional name the questions can hardly be answered since one has to index the fictional name. The new pluri-propositionalist framework focuses on utterances instead of sentences. Rather than being individuated by its referent, 'Sherlock Holmes' is individuated by the fictive narrator, the time and location of the indexed token: e.g., 'Sherlock Holmes' <Watson, end of 19th century, London >.

'Sherlock Holmes' is a fictional name possessing a fictional index that it carries with it in utterances of the name in ordinary conversation, e.g., conversations where (1) is uttered. A speaker's knowledge of the name is also echoed in their understanding of the truth conditions of an utterance such as (1). In the case of (1), we may have something like

> Given that (1) is an English sentence, the utterance **u** of (1) is true if and only if the individual that the convention exploited by the utterance **u** allows us to designate by 'Sherlock Holmes' (Watson, end of 19th century, London) exists.

The truth conditions of this utterance give its cognitive significance. A speaker needs nothing more than these truth conditions to identify the cognitive significance. In the truth conditions associated with an utterance of (1), 'the individual that the convention exploited by the utterance **u** allows us to designate by 'Sherlock Holmes"—where the name is indexed—echoes the fact that 'Sherlock Holmes' is a fictional name.[14]

[14] The cognitive significance containing a name does not make it about that name, and it does not make it metalinguistic.

I know that 'Sherlock Holmes' in (1) is a fictional name, and that the truth conditions mentioned give the truth conditions of the utterance of (1). Not surprisingly, an utterance of (1) has a descriptive content, which is captured in the truth conditions of the utterance. An utterance of (1) can be accepted as true, but it is not true given facts. Suppose that you believe that Sherlock Holmes is a real individual. Then you may proceed to assign an utterance of (1) different truth conditions.

> Given that (1) is an English sentence, the utterance **u** of (1) is true if and only if the individual that the convention exploited by the utterance **u** allows us to designate by 'Sherlock Holmes' exists.

You and I may assign different truth conditions to an utterance of (1), and the utterance also has different cognitive significance for you and I. Moreover, I will not look for a designational content, whereas you may. There is no need to invoke a fictional character to account for our disagreement. In addition, the latter is purely cognitive. It does not, and it cannot, invoke a designational content containing Sherlock Holmes, as a real or even as a fictional creature. In any case, invoking a fictional creature is not required in order to account for the facts. The identification of a specific 'Sherlock Holmes', for instance, is often done indirectly, by means of the name of the author, the title of the book, the title of the movie or the television series, the name of the actor playing Holmes, etc. These are not fiction-relative parameters. Questions pertaining specifically to names can then be asked again: is the name in Doyle's book and the name in a television series the same name? Readers of *A Study in Scarlet* assume that every token of 'Sherlock Holmes' in the book has the same fictional index. Readers of the other books of Doyle's featuring Holmes assume that tokens of 'Sherlock Holmes' in these books have the same or similar indexes. In saying this, I am going beyond fictional names and I am talking about the readers of the fiction. Clearly, this is not part of my view on fictional names.

Moreover, there is no convention tying a fictional name to a three-dimensional individual. An informed speaker using a fictional name knows that there is no such convention, and knows the cognitive significance of an utterance of (1). Such a speaker also knows the difference between the truth conditions as cognitive significance of an utterance involving ordinary names, and the truth conditions as cognitive significance of an utterance involving fictional names. I'll let the reader give the truth conditions of an utterance of (2). There is no need to explore whether or not they are the same name in (1) and (2) here. In any case, the answer will not depend on facts. Accepting as true a fictional indexed token, or sentences containing fictional names, also allows us to set aside the famous prefix or complex operator 'It is true in the story' and any reference to a story.

The fictional name itself is part of the cognitive significance of an utterance of the sentence. If a speaker does not assign an utterance of (1) such truth conditions, then that speaker does not know that the name

is fictional. If the fictional name is poorly individuated, i.e., if it is not assigned a clear index, then the utterance lacks clear truth conditions and its cognitive significance is unclear. I contend that utterances containing fictional names carry no truth conditions confronted with facts. The initial questions concerning existence can only be about semantically determined content or the cognitive significance of the utterance. An utterance of (1), or (2), can be accepted as true, but this assumes or implies nothing concerning facts, i.e., it is not true relative to facts. If the utterance is about designational content, then it is assumed that there is a designational content and that the name refers to an object. The fictional name is then wrongly seen as having the same function as an ordinary name. If we accept utterances of sentences containing fictional names such as 'Sherlock Holmes exists' or 'Sherlock Holmes does not exist' as true, it is not because we have assessed it as true considering facts, since on the account developed here such a suggestion is altogether incoherent. Puzzles concerning the existence, or non-existence, of Sherlock Holmes concerned with facts are grounded on a reading of utterances such as (1) and (2) which misleadingly focuses on the purported designational content of the utterances, while there is none—when instead the emphasis should be the cognitive significance of such utterances.

Reasons to accept or reject an utterance of (1) can differ widely. Determining whether an utterance such as 'Obama exists' expresses a true or false proposition is rather straightforward, since we simply have to check the facts. However, determining whether utterances of (1) and (2) determine proposition one accepts as true, or reject as false, is more complicated, since it relies on things such as the name's individuation grounded on non-linguistic knowledge, knowledge of the name as fictional, the use of an index for that name, views on fiction, and so on.

So far, I have addressed the truth-value problem. The contradiction problem is a little more complicated but it is rather easy to deal with on my account. An utterance of

(5) Sherlock Holmes exists and Sherlock Holmes does not exist

is intuitively a contradiction if the same fictional name (i.e., a name with the same fictional index) occurs in both elementary sentences. One can say that it is a contradiction, not because of facts, but because *prima facie* both sentences cannot be accepted as true by rational speaker. However, it can be accepted as a true contradiction if, for whatever reasons, an utterance of an identity sentence with the first name with a fictional index and the second name with the same fictional index is accepted as true. In any case, assessing an utterance of (5) as a contradiction is grounded on linguistic competence only.

We are left with the belief attribution problem. Consider an utterance of

(6) Donald Trump believes that Sherlock Holmes exists.

An utterance of (6) can be assigned truth conditions. Fixing the referent of 'Donald Trump', we have, with 'Sherlock Holmes' as a fictional name:

> Given that (6) is an English sentence, the utterance **u** of (6) is true if and only if DONALD TRUMP believes that what the convention exploited by the utterance allows us to designate by 'Sherlock Holmes' (Watson, end of 19th century, London) exists.

The belief attributed is fully descriptive. He might also believe that Holmes is real. We then have:

> Given that (6) is an English sentence, the utterance **u** of (6) is true if and only if DONALD TRUMP believes that what the convention exploited by the utterance allows us to designate by 'Sherlock Holmes' exists.

where 'Sherlock Holmes' is, wrongly, believed to be an ordinary name. These are attributions of different beliefs. Of course, Trump cannot have a belief about a singular proposition containing Holmes himself.

We, therefore, have two options—corresponding to one reading where 'Sherlock Holmes' is an ordinary referring name, and another where it is a fictional name. It is interesting to be able to capture these two options and to make them clear in the truth conditions, as cognitive significance, assigned to utterances. We can also have

(7) Donald Trump believes that Sherlock Holmes does not exist

with 'Sherlock Holmes' as a fictional name. An utterance of (7) can be assigned truth conditions

> Given that (7) is an English sentence, the utterance **u** of (7) is true if and only if DONALD TRUMP believes that what the convention exploited by the utterance allows us to designate by 'Sherlock Holmes' (Watson, end of 19th century, London) does not exist.

Finally, it is also possible that the same fictional name as used in (7) is also used in (8)

(8) Obama believes that Sherlock Holmes does not exist.

We can assign an utterance of (8) the following truth conditions and a descriptive content:

> Given that (8) is an English sentence, the utterance **u** of (8) is true if and only if OBAMA believes that what the convention exploited by the utterance allows us to designate by 'Sherlock Holmes' (Watson, end of 19th century, London) does not exist.

In no case is a belief in a singular proposition attributed. Trump and Obama can be attributed the same belief about 'Sherlock Holmes', such a belief being about the cognitive significance of an utterance. However, (7) and (8) might also attribute different beliefs—(7) might contain the name found in Doyle's books and (8) the name found in a television series.

7. *Conclusion*

Mono-propositionalism, based on Frege's ground breaking introduction of senses in semantics or based on singular propositions, is not a framework fit to account for fictional names and the truth conditions of sentences containing fictional names. Pluri-propositionalism offers a new perspective on such sentences. Is an utterance of (9)

(9) Sherlock Holmes is Sherlock Holmes

where the first occurrence of the name is about the 19th century detective—'Sherlock Holmes' with an index containing Watson, end of 19th century and London,—and the second about the 21st century detective—with an index containing Watson, the 21st century and London—true? The question as to whether the first Watson is the second Watson remains open here. One must make a distinction between the cognitive significance of the utterance and its designational content. The utterance of (9) has no designational content and it is not truth assessable relative to the facts. Unfortunately, or fortunately, reasons to accept or reject the utterance of (9) will remain forever an object of speculation. The awkwardness of an utterance of 'Frank Serpico is a detective, and so is Sherlock Holmes' is made clear once the cognitive significance of the utterance is considered. An utterance **u** is true, if and only if, the individual that the convention exploited by the utterance **u** allows us to designate by 'Franck Serpico' is a detective, and also if and only if, the individual that the convention exploited by the utterance **u** allows us to designate by 'Sherlock Holmes' (Watson, end of 19th century, London) is also a detective. The first sentence determines a designational content containing an individual, whereas the second does not.

Fictional names do not refer to fictional or possible objects. If I am right, a fictional name is not a variable assigned a value, and it cannot be modeled along these lines: there simply is neither an assignment nor a value. A fictional name is a sequence of letters, which is indexed to a fiction, with a fictional narrator, a time, and a location. This alternative picture opens up new perspectives on fictional names and, in a sense, on fiction itself. Understanding sentences containing fictional names is a purely cognitive affair and it does not require invoking fictional entities. I do not wish to deny the ontological problems raised by fiction or to disqualify an examination of fictional creatures (see Thomasson 1990, Voltolini 2006, Sainsbury 2014, Kripke 2013). Whatever these problems, they have no impact on my view, and, I submit, on the semantics of fictional names. It also suggests they do not depend on fictional names in stories. By the same token, fictional names have no impact on ontological issues concerning fiction.

Acknowledgements

I would like to thank Paul Bernier and Dylan Hurry for very helpful comments on a previous version of this paper.

References

Adams, F. 2011. "Sweet Nothings: The Semantics, Pragmatics, and Ontology of Fiction." In F. Lihoreau 2011: 119–135.

Braun, D. 1993. "Empty Names." *Noûs* 27 (4): 449–469.

Braun, D. 2005. "Names, Fictional Names, and Mythical Names." *Noûs* 39 (4): 596–631.

Currie, G. 1990. *The Nature of Fiction*. New York: Cambridge University Press.

Everett, A. and T. Hofweber (eds.). 2000. *Empty Names, Fiction and the Puzzles of Non-Existence*. Stanford: CSLI publications.

Everett, A. 2003. "Empty Names and 'Gappy' Propositions." *Philosophical Studies* 116: 1–36.

Everett, A. 2007. "Pretense, Existence, and Fictional Objects." *Philosophy and Phenomenological Research* 74 (1): 56–80.

Korta, K. and J. Perry 2011. *Critical Pragmatics*. Cambridge: Cambridge University Press.

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

Kripke, S. 2011. "Vacuous Names and Fictional Entities." In *Collected Papers 1*, Oxford: Oxford University Press.

Kripke, S. 2013. *Reference and Existence*. Oxford: Oxford University Press.

Kroon, F. 2014. "Content Relativism and the Problem of Empty Names." In M. Garcia-Carpintero and G. Marti (eds.). *Empty Representations. Reference and Non-Existence*. Oxford: Oxford University Press 2014: 14–164.

Lihoreau, F. (ed.) 2011. *Truth in Fiction*. Frankfurt: Ontos Verlag.

Perry, J. 2012. *Reference and Reflexivity*. 2nd edition. Stanford: CSLI Publications.

Predelli, S. 2002. "'Holmes' and Holmes – A Millian Analysis of Names of Fiction." *Dialectica* 56 (3): 261–279.

Sainsbury, M. 2011. "Fiction and Acceptance-RelativeTruth, Belief and Assertion." In Lihoreau 2011: 119–135.

Sainsbury, M. 2014. *Reference Without Referent*. Oxford University Press.

Salmon, N. 1998. "Nonexistence." *Noûs* 32 (3): 277–319.

Thomasson, A. 1999. *Fiction and Metaphysics*. New York: Cambridge University Press.

Thomasson, A. 2003. "Speaking of Fictional Characters." *Dialectica* 57 (2): 205–223.

Vallée, R. 2018. "Fictional Names and Truth." *Organon F* 25 (1): 74–99.

Voltolini, A. 2006. *How Ficta Follow Fiction. A syncretistic Account of Fictional Entities. Dordrecht*. New York: Springer.

Walton, K. 1990. *Mimesis as Make-Belief*. Cambridge: Harvard University Press.

Walton, K. 2000. "Existence as Metaphor?" In A. Everett and T. Hofweber (eds.). *Empty Names, Fiction and the Puzzles of Non-Existence*. Stanford: CSLI publications: 69–94.

Croatian Journal of Philosophy
Vol. XX, No. 58, 2020

Epistemic Infinitism, the Reason-Giving Game, and the Regress Skeptic

ERHAN DEMIRCIOĞLU
Koç University, Istanbul, Turkey

Epistemic infinitism is one of the logically possible responses to the epistemic regress problem, claiming that the justification of a given proposition requires an infinite and non-circular structure of reasons. In this paper, I will examine the dialectic between the epistemic infinitist and the regress skeptic, the sort of skeptic that bases his attack to the possibility of justification on the regress of reasons. I aim to show that what makes epistemic infinitism appear as well-equipped to silence the regress skeptic is the very same thing that renders it susceptible to a powerful skeptical assault by the regress skeptic.

Keywords: Epistemic infinitism, the epistemic regress problem, skepticism, inferential justification, Peter Klein.

But who will guard the guardians?
Juvenal, The Satires 6.029–34

1. *Introduction*

What are the conditions a given set of beliefs must meet in order for those beliefs to have some positive epistemic status such as being justified or reasonable? Epistemological theories that attempt to answer a question of this sort might be plausibly called "normative". Do beliefs of the sort typically held by human beings actually meet the conditions they must meet in order for them to have the desired positive epistemic status? Epistemological theories that attempt to answer a question of this sort might be plausibly called "descriptive". I simply introduce "normative" and "descriptive" as labels, hopefully in a way that reflects a clear sense that might be plausibly attached to them, but without reading too much into them.

The main epistemological theories about the structural conditions a given set of beliefs must meet in order for them to be justified might be conceived either as being merely normative or as being both normative

and descriptive. For instance, foundationalism conceived as a merely normative epistemological theory claims, roughly, that a given set of beliefs must be structured like a pyramid if they are to be justified, at the bottom of which there are justified foundational beliefs and at the upper layers of which there are beliefs that are justified ultimately by their evidential relations to the foundational beliefs. And, foundationalism conceived as both a normative and a descriptive theory claims, roughly, not only that a given set of beliefs must be structured like a pyramid but also that beliefs of the sort typically held by human beings are structured like a pyramid. So, foundationalism qua a normative and descriptive theory in the sense at issue is *eo ipso* anti-skeptical, but foundationalism qua a merely normative theory need not be. A skeptic might approach foundationalism qua a merely normative theory in three broadly different ways. One is to argue that foundationalism cannot be and therefore is not the correct normative theory (for instance, because, the skeptic might say, the notion of justified foundational belief is incoherent)—let us call this "the normative skeptical approach." Another is to argue that whether or not the structural norms prescribed by foundationalism are correct, there is something about those norms that entails that no beliefs can be justified (for instance, the skeptic might argue that justification can only be "internal" and yet that foundational beliefs can only be justified if "external" justification is possible)—let us call this "the skeptical outcome approach." Yet another one is to argue that beliefs of the sort typically held by human beings do not rise to the challenge of satisfying the structural norms prescribed by foundationalism, whether or not those norms are correct—let us call this "the descriptive skeptical approach." Of course, these three skeptical approaches are not mutually exclusive.

Foundationalism is the oldest game in the town of epistemologists. In this paper, I am interested in a relatively novel contender, namely, infinitism qua a merely normative theory, an epistemological theory whose history is characterized either by comforting oblivion or by quick dismissal but that has received considerable and well-deserved attention in recent years mainly due to the pioneering defense of Peter Klein.[1] More specifically, I am interested in whether the regress skeptic can plausibly adopt what I have labelled "the skeptical outcome approach" against the infinitist.[2] As we shall see, there are good reasons, stem-

[1] Here is a representative but incomplete list of Klein's works that defend epistemic infinitism: (1998, 1999, 2000, 2003, 2005a, b, 2007a, b, 2011, 2014). Klein's works have inspired a large and still growing literature on infinitism. Here is again a representative but incomplete sample: Fantl (2003), Cling (2004), Aikin (2005, 2008, 2011), Wright (2013).

[2] A formidable and familiar skeptical challenge against infinitism is descriptive, taking its cue either from the finiteness of the human mind or the finiteness of the amount of time available to human beings. In this paper, I leave the descriptive skeptical approach aside, which is not to say that the apparent contrast between what infinitism demands and what human beings can in principle offer is of no

ming from the norms governing what one might call "the reason-giving game," for thinking that infinitism is tailor-made for defusing the skeptical outcome approach and it turns out that it is unclear whether the regress skeptic is in a position to adopt that approach.

This paper is hereafter divided into seven sections. In section 2, I attempt to show how infinitism is naturally suggested as the correct normative account of epistemic justification by the reason-giving game. In section 3, the move from the reason-giving game to infinitism is clarified by making some of its central assumptions explicit and is defended against some main objections. In section 4, I raise the question of whether the skeptic is in a position to adopt the skeptical outcome approach against infinitism, given that the reason-giving game between the skeptic and a given subject meeting the norms of infinitism ought rationally to result in a tie. A proper answer to this question, I maintain, requires taking a closer look at the question of why exactly infinitism is suggested by the reason-giving game, and I argue in section 5 that the answer to the latter question lies in a particular feature of inferential justification. In section 6, I claim that the very feature of inferential justification I specify that is responsible for why infinitism is suggested by the reason-giving game can be deployed by the regress skeptic in an argument against infinitism: an ultimate tragedy of infinitism is that what makes it appear as well-equipped to silence the skeptical outcome approach is the very same thing that renders it susceptible to a powerful assault by that approach. Section 7 discusses some infinitist responses to the skeptical assault and finds them inconclusive. Section 8 sums up the lesson.

2. *The reason-giving game and epistemic infinitism*

Michael and Susan are two intellectually sophisticated subjects and they decide to play a game, one quite familiar to epistemologists, which they call "the reason-giving game." They pick a proposition, P, and they both assume that Michael believes that P. Now, Susan is the "detective", and adopts the role of an inquisitive and "maximally persistent"[3] inquirer whose aim is to discover whether Michael's belief that P is epistemically justified. And, Michael is the "defender", as persistent as the detective, whose aim is to persuade Susan that his belief that P is epistemically justified by defending it. The game comes in "steps",

epistemological significance. The normative skeptical approach against infinitism is less popular but formidable all the same. The skeptic might argue, for instance, that infinitism requires the existence of an actual infinity and also argue, along with Aristotle, that the notion of actual infinity is incoherent—hence that infinitism cannot be the correct account of norms governing justification. In this paper, I also leave the normative skeptical approach aside.

[3] Compare Leite (2005): "Imagine that someone invites you to defend a belief. You offer what you take to be a good reason for believing as you do, but your interlocutor asks you to support this reason and continues in like fashion in each step…I call this character 'the persistent interlocutor'" (p. 397).

composed of a "what reason?" question and a "my reason" answer. The first step starts with a question that Susan raises, "What reason do you have for believing that P?", and ends with Michael's citing reasons that support P, reasons which they both assume Michael believes and thus sincerely offers. They both agree that a reason cited in support of P does not render Michael's belief that P justified if that reason itself needs to be supported but is not supported by further reasons. So, the game does not end at the first step if the reason Michael cites in support of P needs to be supported in order for it to justify his belief that P. Susan's job is to reiterate the question at each step in a suitable form, and Michael's job is to answer it by citing new reasons. They both agree that Michael loses the game if all he can do at a particular step is to cite a reason which he has cited at a previous step or if there is a step at which he cannot cite any new reasons. They also agree that Michael wins the game if, and only if, there comes a step where Susan ought to be persuaded that Michael's belief that P is epistemically justified. And, by this, they mean that Michael wins the game if, and only if, there comes a step at which the "what reason?" question cannot be legitimately iterated, i.e., cannot be iterated without its losing the rationale it serves at the very first and previous steps, the rationale in virtue of which the game has kept going until that step.

A natural and tempting line of thought delivers the result that Michael cannot win the reason-giving game conceived in the way above.[4] The rationale of raising the "what reason?" question at the first step is that Michael's persuading Susan that his belief that P is justified requires an answer to that question. For all Susan knows at the outset, Michael's belief that P might be based on a sheer guess, a hunch, or might simply have come "out of the blue", in which case she ought not to be persuaded that it is a justified belief. As the first step ends with Michael's citing a reason, R_1, for P, the rationale of raising the "what reason?" reason remains intact at the second step: for all Susan knows at the second step, Michael's belief that R_1 might be based on a sheer guess, a hunch, or might simply have come "out of the blue", in which case she ought not to be persuaded that his belief that R_1 or his belief that P is justified. And, obviously, after the second step ends with Michael's citing a reason, R_2, for R_1, the rationale of raising the "what reason?" question is kept at the third step, and the same goes for all the subsequent steps. This means that, whatever n is, the reason R_n Michael cites at the n^{th} step needs to be supported in order for him to persuade Susan that his belief that R_n and his belief that P is justified. So, there is no step at which the "what reason?" question can possibly lose the rationale it has at the previous steps, which entails that Michael cannot win the reason-giving game.

[4] That Michael cannot win the reason-giving game above does not follow from the way the game is defined above. (Thanks to a reviewer for pressing on this point.) An argument that Michael cannot win the reason-giving game is offered in what follows.

It is thus plausible to say that Michael cannot win the reason-giving game and that the best he can do is *not* to lose it. In order for him not to lose the game, Michael must be in a position to cite a reason at each step, one that he has not cited at one of the previous steps; and, since the "what reason?" question can (in principle) be legitimately iterated by the maximally persistent inquirer Susan indefinitely, the following must be true of Michael if he is not to lose the game: there must be an infinite set of reasons available to Michael arranged in a non-repeating series such that the first member, R_1 is a reason for P, and the second member, R_2 is a reason for R_1, and the third member, R_3 is a reason for R_2, and so on. So, in order for Michael not to lose the game, the structure of his reasons must be infinite and non-repeating. That is to say, his reasons must be structured in the way epistemic infinitism says they must. Epistemic infinitism is a normative epistemological thesis about how our beliefs must be organized in order for them to be justified, and it claims that a necessary condition for a series of reasons to lend justification to a proposition is that it must have no repeating members and has no last member. A general moral that can be plausibly drawn from the reason-giving game is that the best we can do in the way of having justified beliefs is by having at our disposal an infinite chain of reasons structured in the way the epistemic infinitist says it must.[5]

An important distinction any account of epistemic justification including epistemic infinitism must respect is between propositional and doxastic justification.[6] Propositional justification is a property of a proposition that it has relative to a given subject, irrespective of whether the subject believes the proposition. We can say that a *proposition*, P, is propositionally justified for a subject, S, only if there is a (good) reason for P that is available to S (irrespective of whether S believes that P). Doxastic justification, on the other hand, is a property of a subject's already formed believing attitude towards a certain proposition. We can say that S's (actual) *believing* that P is doxastically justified just in case P is propositionally justified for S and S bases his believing attitude on the reason (or the chain of reasons) by which P is propositionally justified for S. Doxastic justification is thus propositional justification plus the basing relation.[7] This is a tidy picture, and here is a little complication: suppose S believes that P, and P is justified for S, but S's believing that P is not justified (because the basing requirement is not satisfied). What can we say about the justificatory status of S's *belief* that P—is it justified or not? Suppose one simply insists for a "yes-or-no" answer (and does not accept a "yes-and-no" answer). Clearly, such an answer to this question must be stipulative in nature. I hereby make the stipula-

[5] Compare Aikin: "In essence, the thought behind the (infinitist) view is that if you know, you can answer questions about what you know until there just aren't any more questions. But, as it turns out, there are in principle no final questions. So, knowers need to be able to keep coming with answers" (2009: 57).

[6] The distinction is first introduced by Firth (1978).

[7] For a critical discussion, see Turri (2010).

tion that S's belief that P is justified if and only if S believes that P and P is propositionally justified for S. Given this, the answer I give to the question is yes. In what follows, whenever I talk about the justificatory status of a given *belief* without qualification, I will have in mind solely considerations that pertain to the propositional justification of its content.

Now, the question that arises is this: is Susan the detective attempting to figure out whether (i) Michael's *believing* that P is justified or whether (ii) Michael's *belief* that P is justified? The answer is "both". This is because what Susan is at bottom interested in is whether P is propositionally justified for Michael, and also because both Michael's believing that P and Michael's belief that P requires that P be propositionally justified for Michael. To see why, consider the following. If one opts for (i), then Susan is to be conceived as trying to discover whether Michael's believing that P is doxastically justified. Since doxastic justification requires propositional justification and basing, Susan is then trying to discover whether P is propositionally justified for Michael and Michael bases his believing attitude towards P on the reason by which P is propositionally justified for Michael. Furthermore, since it might be plausibly assumed that a subject bases his believing attitude towards a proposition on a reason if (though not necessarily *only* if) he sincerely cites the reason in support of the proposition, then the basing requirement for doxastic justification is *automatically* satisfied as soon as Michael sincerely cites a reason in response to a "what reason?" question raised by Susan at a particular step. So, if one opts for (i), then one is entitled to say that what Susan is at bottom trying to discover is whether P is propositionally justified for Michael. And, if one opts for (ii), Susan is to be conceived as trying to discover whether P is propositionally justified for Michael and Michael believes that P. Since it is assumed by both of our subjects that Michael believes that P (and the reason he offers in support of P and the reason he offers in support of the reason he has offered for P, and so on), one is entitled to say, if one opts for (ii), that what Susan is at bottom trying to discover is whether P is propositionally justified for Michael. So, irrespective of whether (i) or (ii) is to be adopted, the reason-giving game between Michael and Susan centers on the question of whether P is propositionally justified for Michael. Accordingly, epistemic infinitism that is strongly suggested by the reason-giving game is to be conceived, at least in the first instance, as an account of the condition a subject must meet in order for a proposition to be *propositionally* justified for him. Epistemic infinitism about propositional justification is the claim that the condition a subject must meet in order for a proposition to be justified for him is that there must be infinitely many reasons available to him structured in a non-repeating way.

3. *Caveats*

Various clarifications and qualifications are required in order to fortify the move from what we can plausibly derive from the reason-giving game to epistemic infinitism. First, a main moral of the reason-giving game is that there is no forthcoming step at which the "what reason?" question loses the rationale it has at the previous steps. This is consistent with the fact that at times we seem to be engaging in something relevantly like the reason-giving game and adopting the detective role in ordinary quotidian contexts, there is always an n^{th} step at which we qua ordinary beings with finite amount of time, perseverance, and guided mainly by pragmatic concerns *concede* a reason provided with us at that step (or perhaps resolve the issue with our fists or just leave the scene). The idea, however, is that we are never *rationally* compelled to do so, and as such the moral of the reason-giving game is normative and abstracts away from what we actually do or tend to do in similar circumstances.[8]

Secondly, and relatedly, the rules of the reason-giving game and the rationale behind it appear to *isolate* some of the core features of the *ordinary* conception of how a rational dialectic between two speakers should go, rather than resting on or taking for granted a set of standards that are far removed from the standards governing an ordinary dialogue that we would ordinarily take to be rational. There are surely everyday conversational contexts in which a particular speaker might wish to figure out the reason behind one of the beliefs of the other speaker. "So, you believe that Trump will make America great again, why is that?" A question like this might normally stem from a suspicion about the truth of what is believed or from a desire to see whether the person that has the belief is justified. In any case, we do not feel that the question is always "off the mark", "odd" or "inappropriate" but can easily imagine everyday contexts in which the person that has the belief *ought* to provide an answer if his belief is to count as reasonable or justified. Now, it is true that ordinary conversational contexts are typically characterized by a "common ground," a set of "background assumptions" that are shared by speakers and ultimately serve as dialectical regress-stoppers. Beliefs about basic arithmetic ("2+2=4"), the

[8] A reviewer raises the following worry: "If I assert that I ate potatoes for breakfast, and someone says I'm not reasonable in believing unless I can publically defend it, I think they are wrong. And, I think it has nothing to do with how much time I have on my hands." In response, let me first note that I claim above that the fact that there is no forthcoming step in the reason-giving game at which the "what reason?" question loses the rationale it has at the previous steps is consistent with the fact that there is always an n^{th} step at which we actually concede a reason provided with us at that step. I have argued in the previous section for the claim that there is no forthcoming step at which the "what reason?" question loses its initial rationale. So, if the reviewer cannot answer the "what reason?" question raised for the belief that I ate potatoes for breakfast, then s/he would *lose* the reason-giving game as the game is *defined*. What this has to do with *justification* is clarified below.

immediate environment ("Here is a hand"), one's own occurrent mental states ("I feel pain"), "hinge" propositions ("There is an external world") do usually provide the background against which ordinary dialectical exchanges take place. So, there is a sense in which an attempt to question what is ordinarily taken to be a "common ground" between speakers is bound to appear "off the mark" or "odd". However, despite this, it might be plausibly argued that there is again no special difficulty ordinary speakers feel in admitting that there is a clear sense in which just as any other belief, the beliefs belonging to the common ground are not unquestionable but stand in need of the support of reasons, and that the fact that our interlocutors usually let us get away with making such assertions as "I have a hand" is consistent with the fact that we as ordinary speakers can easily imagine conversational contexts in which questioning them is appropriate. Why doesn't the idea that the beliefs ordinarily viewed as belonging to the common ground are not unquestionable strike ordinary speakers as strange, wild, or absurd, as something that they ought to reject given their conception of a rational conversation? It is, it might be argued, because there is a norm ordinary speakers are willing to admit that applies to all beliefs across the board: be ready to provide reasons for any of your beliefs when challenged, if they are to count as reasonable. The reason-giving game takes for granted and is built on a norm along these lines, a norm, one might plausibly argue, that is not "strange" or "wild" but is treated by ordinary speakers as ultimately correct.[9]

Thirdly, and relatedly again, the argument that Michael cannot win the reason-giving game presupposes what one might call the "unrestricted-defense" view, according to which all beliefs asserted in the reason-giving game require defense in the light of requests for reasons—when challenged by the "what reason?" question. On this view, there are no privileged beliefs whose assertions cannot be legitimately disputed by raising the "what reason?" question. The unrestricted-defense view can be contrasted with what one might call the "unrestricted-challenge" view and the "restricted-challenge" view. The unrestricted-challenge view holds that *all* beliefs are "presumptively rational" or are "defeasible presumptions" in that one can only challenge their assertions by providing grounds or reasons for doubting them and not by merely raising the "what reason?" question.[10] The restricted-challenge view holds that only some beliefs are presumptively rational.[11] I will not attempt to adjudicate between these views, but remain content with maintain-

[9] The central aim of this section is to disclose those main assumptions that connect the reason-giving game to epistemic infinitism, while showing that those assumptions are not gratuitous. The argument provided above is not, and is not intended to be, a decisive argument on behalf of the norm in question, but it is, and is intended to be, an argument that shows that the norm is to be taken seriously. Thanks to a reviewer's comment that prompts this note.

[10] See Adler (2002).

[11] See Brandom (1994).

ing that the argument that Michael cannot win the reason-giving game presupposes the unrestricted-defense view.

Fourthly, the reason-giving game presupposes that a given subject's belief that P is justified only if the subject has the ability to defend (or is in a position to cite a reason for) the belief. This is questionable. It seems clear that non-linguistic creatures and human infants can have justified beliefs despite the fact that they are not able to cite reasons for their beliefs. Moreover, adult humans may have justified beliefs despite the fact that the original reasons for those beliefs, though compelling, have long since been forgotten. The fact that the subject is *now* at a loss if asked to justify his belief does not show that his belief is thereby unjustified. The main point here is that the state of holding a justified belief is to be distinguished from the activity of justifying a belief or from having the ability to justify it. This is an important point, which must be granted, and it calls for a qualification. The qualification required is this: there is *a sense* of justified belief in which a subject's belief that P is justified only if the subject has the ability to defend (or is in a position to cite a reason for) the belief. This is the sense of justified belief as essentially the product of the reflective activity of examining our beliefs and determining which, if any, are worthy of being kept. And, this is the sense of justified belief that accords well with the notion of a responsible epistemic agent seeking to retain only those beliefs worthy of being retained.[12] In *this* sense of the term, it seems that the distinction between the state of holding a justified belief and the activity of justifying a belief is largely bogus. And in this sense of the term, neither non-linguistic creatures nor human infants can have justified beliefs. And the same goes for adult humans that have forgotten the reasons they once had for their beliefs.

Fifthly, and finally, the move from the fact that the best we can do in the reason-giving game is by having at our disposal an infinite and non-repeating chain of reasons to *epistemic* infinitism is suspicious. Here it must be observed that the reason-giving game involves a discursive practice, a dialectical interaction between two subjects. And, epistemic infinitism is a thesis about how the propositions available to a subject must be structured in order for the subject to be justified in believing those propositions. Now, it is questionable whether the normative rules governing a rational discursive practice have anything essential to do with the epistemological concerns about the structure of justified beliefs. In particular, it might be claimed that while it might be true that in order for Michael to be rational in his attempt to *defend* his belief that P or to *persuade* Susan that his belief that P is justified, he must be in a position to cite new reasons at each step, it does not follow that Michael's belief that P is justified only if the *structure* of his

[12] Compare also Aikin: "Epistemic infinitism…holds that those who know are those who have been *maximally intellectually responsible*…Who would say that someone knows that p, if asked why he believes it, he shrugged his shoulders and uttered an inarticulate "hmmm… idunno"?" (2009: 57–8, emphasis mine).

reasons must be infinite and non-repeating. So, it might be argued that one can, for instance, consistently defend epistemic foundationalism (the view that the epistemic regress must halt at basic beliefs (i.e. justified beliefs that are not justified in virtue of other beliefs) if one is to have justified beliefs at all) while acknowledging that rational defense or persuasion in the reason-giving game requires having the capacity to cite new reasons at all forthcoming steps: epistemic infinitism does not follow from what one might call dialectical infinitism. This is an important point but there are two things that can be said in response. One is that the reason-giving game does not strictly require two subjects and can be played with only one subject adopting both the roles of a detective and a defender. If the moral of the reason-giving game with two subjects is dialectical infinitism, then the moral of that game with only one subject engaged in a *sotto voce* dialog is also dialectical infinitism. The other is that the distinction between epistemic and dialectical infinitism is again largely bogus if the relevant sense of justified belief is one that conceives justification as an epistemic status that the subject must earn through engaging in an activity of justifying in order for him to have justified beliefs, that is, if having justification rests essentially on an activity of justifying.[13]

With these points in mind, it appears that the move from what must be true of Michael, our hypothetical subject defending his belief that P, in order for him not to lose the reason-giving game to epistemic infinitism is safe. In this connection, it is important to realize that Klein, a prominent epistemic infinitist, actually argues against epistemic foundationalism by an argument that rests on the normative rules governing reason-giving procedures (2004: 14–15). Consider an epistemic foundationalist, Fred, who takes his belief that P to be epistemically basic. Sally, a persistent interlocutor, asks Fred his reason for believing that P. According to Klein, Fred faces a dilemma here. He may either simply say that there is no reason that he can offer for believing that P but still insist that P, or realize that it is in virtue of its having a certain property, F, he takes the belief that P as epistemically basic and say that the belief that P has F and that beliefs with F are likely to be true. If Fred takes the first horn, then his belief that P is dogmatic and he ought rationally to abandon it. And, if he takes the second horn, then the regress continues by the question "What reason do you have for thinking that beliefs that have F are likely to be true?" and *contra* Fred the epistemic foundationalist, his belief that P is not a regress-stopper and is thus not epistemically basic. The lesson Klein derives about *the reason-giving game* Fred and Sally engage in is that the epistemic foundationalist "can't be an epistemically responsible agent and practice what he preaches" (2004: 15). And, this lesson about the reason-giving game is what grounds Klein's claim that *epistemic foundationalism* "advocates accepting an arbitrary reason at the base" (1999: 297).

[13] For further discussion, see Rescorla (2009).

4. Two virtues? Or just one?

If our beliefs are structured in the way epistemic infinitism says they must in order to be justified, then it seems that we are in a position to alleviate a major skeptical worry that might arise about the justificatory status of our beliefs. This is clearly so, given that one of the favorite tools the skeptic uses to question the justificatory status of our beliefs is the very same question the detective raises in the reason-giving game, viz. the "what reason?" question. Let us call the philosopher that bases her skeptical attack to the possibility of justification on the regress of reasons *the regress skeptic*. Now, suppose that the reasons available to me for believing in a particular proposition are infinitely many and non-repeating, which means that I am in a position to cite a reason for each reason that I offer and might be challenged by the regress skeptic. If this is so, then I cannot lose, and the regress skeptic cannot win, the reason-giving game. It is also true that I cannot win, and the regress skeptic cannot lose, it. However, it might be argued that not losing the reason-giving game against the regress skeptic, our notoriously powerful opponent, is perhaps victory *enough*.

The "what reason?" question raised for a particular belief starts a regress of reasons. And, if we maintain that there is no forthcoming step at which the "what question?" loses the rationale it has at the previous steps, then the only way for us not to lose the reason-giving game is by having at our disposal an infinite (and non-repeating) chain of reasons. Epistemic infinitism is simply the view that fully endorses the regress of reasons strongly suggested by the reason-giving game: it appears to be a direct outcome of an appreciation of a natural (if not uncontroversial) conception of the rules governing the reason-giving game. Furthermore, it seems that it enables a sort of response to the regress skeptic that does not look desperate: if the reasons available to me for a belief is structured in the way epistemic infinitism says it must in order for me to have a justified belief, then I am immune to skeptical challenges to the grounds of that belief in the form of "what reason?" questions.

It thus appears that there are two virtues of epistemic infinitism:

> (RR) Epistemic infinitism takes at face value what is suggested by the reason-giving game: in order for a subject to be justified in believing a proposition, there must be an infinite set of reasons available to the subject arranged in a non-repeating series such that the first member, R_1 is a reason for P, and the second member, R_2 is a reason for R_1, and the third member, R_3 is a reason for R_2, and so on. (RR = Regress is Real)
>
> (RS) If a subject's reasons for a particular belief are infinite in number and structured in the way the epistemic infinitist says they must in order for that belief to be justified, then the reason-giving game concerning that belief with the regress skeptic ought

rationally to result in a tie and therefore skeptical attacks from the regress of justification are circumvented. (RS = Response to the regress Skeptic)

(RR) and (RS) provide strong support for epistemic infinitism. As for (RR), we can say this: it is in general a merit of a theory that it does not yield a gap, or reduces the already-existing gap, between how things "appear" to us (not necessarily in the visual or perceptual sense) and how things really "are". Any theory that yields such an appearance-reality gap faces the often-not-lifted burden of explaining why things "appear" to us differently from how things really "are". It is therefore a virtue of epistemic infinitism that it endorses what "appears" to be a moral of the reason-giving game as a condition for propositional justification. As for (RS), a general point is that any normative epistemological theory that provides an adequate response to the skeptic is preferable to those that do not. This is again because it "appears" to us that we have justified beliefs and we want to resist the skeptical thesis that we have none.

(RR) and (RS) deserve a critical examination. One question that we might ask about (RR) is this: why exactly is epistemic infinitism suggested by the reason-giving game? And, one question that we might ask about (RS) is this: is it true that skeptical attacks from the regress of justification are circumvented, given that the reason-giving game for that belief ought rationally to result in a tie? I will argue that an adequate answer to the question about (RR) paves the way for a "no" answer to the question about (RS). The reason why epistemic infinitism is suggested by the reason-giving game is the reason why the regress skeptic can plausibly argue that the subject is not justified in having a particular belief while appreciating that the reason-giving game for that belief ought rationally to result in a tie.

5. *RR and inferential justification*

Reflecting on (RR), let us start with observing that the reason-giving game presupposes an *argumentative* model of dialectical interaction between the detective and the defender. According to this model, a proper answer by the defender to the "what reason?" question raised by the detective with respect to one of the defender's beliefs requires citing a reason (a proposition the defender believes) that effectively serves as a premise that purportedly supports the belief. Now, assuming that there is a sense of justification (or justified belief) on which an argumentative model of dialectical interaction is also a model of epistemic justification (see section 3), the reason-giving game presupposes an argumentative model of *epistemic justification* (in that sense). Furthermore, since a belief's being inferentially justified is a matter of its owing its justification to the support of other beliefs, an argumentative model of epistemic justification presupposes that only those beliefs

that are inferentially justified are justified. So, given that the reason-giving game presupposes an argumentative model of epistemic justification, epistemic infinitism suggested by that game presupposes that only those beliefs that are inferentially justified are justified.

A crucial point here is that the infinite regress of reasons suggested by the reason-giving game and endorsed by epistemic infinitism has to do with the nature of inferential justification. There are two individually necessary and jointly sufficient conditions for the inferential justification of a belief, captured by the following thesis:

> (IJC) A belief held by a subject is (prima facie) inferentially justified if, and only if, (i) that belief is (adequately) supported by some of the other beliefs of the subject and (ii) those other beliefs of the subject are themselves justified.

Let us call the condition captured by (i) *the support condition*, viz. that in order for a belief to be inferentially justified, there must be another belief that (adequately) supports it (or there must be a suitable 'evidential' relation between the two beliefs). It is notoriously difficult to give an adequate account of the notion of evidential support; but fortunately, I can leave it unanalyzed in this paper. Intuitively, my belief that my wife is back home from the gym is supported by my belief that her car is parked outside but not supported by my belief that Paris is the capital of France. What needs to be observed for the purposes of this paper is that the support thesis is unquestionably true simply because it specifies in part what it means to be inferentially justified. Let us call the condition captured by (ii) *the other-belief-justification condition* (or simply *the justification condition*), viz. that in order for a belief to be inferentially justified by another belief, the latter belief itself must be justified. There is good reason to think that the justification condition is required for inferential justification. Suppose that John believes that his boss is going to fire him and this belief is (adequately) supported solely by one of his other beliefs, viz. that his boss distrusts him. But suppose that the belief that his boss distrusts John is in turn entirely unsupported—John has no reason at all to believe this, and his 'paranoid' tendencies are active in the formation of this belief. In this case, is the belief that John's boss is going to fire him inferentially *justified* by the belief that John's boss distrusts him, given that the former is *supported* by the latter? The answer appears to be a clear "no": the belief that John's boss is going to fire him is not justified and therefore not inferentially justified, and the reason why this is so is evidently that the supporting belief that John's boss distrusts him is not justified. And, this is just to say that the justification condition is not met by John's belief that his boss is going to fire him.

The justification condition for inferential justification generates the regress of reasons in the reason-giving game. Citing a reason, R_1 in support of the target belief that P is not sufficient for the justification of the belief that P: R_1 must also be justified. Without R_1 being justified,

the belief that P is not justified in virtue of its being supported by R_1. So, the initial question of whether the belief that P is justified has not yet been answered by citing R_1. In order to answer that question, we need to know whether R_1 is justified. And, citing R_2 in support of R_1 is not sufficient for the justification of R_1: R_2 must also be justified. Without R_2 being justified, neither R_1 nor the belief that P supported by R_1 is justified in virtue of R_1's being supported by R_2. So, the initial question of whether the belief that P is justified has not yet been answered by citing R_1 in its support and citing R_2 in support of R_1. The same point clearly applies to all the forthcoming steps. It is because the justification condition holds for inferential justification that an answer to the initial question regarding the justificatory status of a target belief requires there being available to the subject an infinite series of reasons.

To further appreciate the connection between the justification condition and the idea that inferential justification requires an infinity of reasons, suppose that (IJC) is rejected in favor (IJS), which reads:

> (IJS) A belief held by a subject is (prima facie) inferentially justified if, and only if, that belief is (adequately) supported by some of the other beliefs of the subject.

(IJS) is what one gets by dropping the justification condition from (IJC). If (IJS) were the principle that is true of inferential justification, then there would be no troubling regress of justification because it would then be sufficient for the inferential justification of a belief that the subject has another belief that evidentially supports it, whether or not that other belief itself is justified. So, if (IJS) were true, John's belief that his boss is going to fire him, for example, would be justified on the basis of the support it gets from his belief that his boss distrusts him. The question about the justificatory status of the target belief (that John's boss is going to fire him) would then be *settled* by John's citing the belief that his boss distrusts him. True, we could still raise the "what reason?" question for the belief that John's boss distrusts him. So, there will still be a *sort* of regress of reasons. But the crucial point is that the sole rationale for raising that question would then be to figure out whether *that* belief itself is justified, *not* whether the original target belief that John's boss is going to fire him is justified. In other words, if (IJS) were true, the belief that John's boss is going to fire him would no longer be 'targeted' by the "what reason?" questions raised in subsequent steps once another belief that adequately supports it is cited in its defense. And, that is what makes the ensuing regress non-troubling against the skeptic questioning the justificatory credentials of our beliefs. If (IJS) were true, then we would *have* as many justified beliefs as the number of our beliefs that receive adequate support from other beliefs we have. This means that if (IJS) were true, epistemic infinitism would not be suggested by the dialectic involved in the reason-giving game as the correct account of justification.

However, given (IJC), the "what reason?" question raised at each step is an attempt to figure out whether John's original belief under scrutiny—the belief that his boss is going to fire him—is justified. Its justificatory status is not settled by citing another belief that adequately supports it as long as the justificatory status of that belief is not settled. And, this is what makes the ensuing regress troubling against the regress skeptic. Given (IJC), the "what reason?" question keeps targeting the very first belief for which it is raised at all forthcoming steps, and this is what suggests that the justification of one and the same belief requires an infinity of reasons, i.e. what suggests epistemic infinitism as the correct normative account of epistemic justification.

Before proceeding further, there is one final point I want to make about how the "if and only if" in (IJC) is to be understood. Suppose that a given subject's belief that P is supported only by one of her other beliefs, R_1. Suppose further that the belief that P is (inferentially) justified. If this is so, then given (IJC), R_1 must be justified. However, R_1's being justified is not merely necessary for P being justified. It is *in virtue of* R_1's being justified that P is justified—R_1's being justified *explains why* P is justified. There is a sort of explanatory dependence relation between P being justified and R_1 being justified, one that is not captured by noting that R_1 being justified is necessary for P being justified. Klein's remarks are helpful here:

> Consider a line AB and some subsegment of it, say s. Now, s is a subsegment of AB only if there is another subsegment of s, say s_1, that is not identical to s (or AB), and there is some subsegment, s_2, etc. In addition, any subsegment *consists* (*in part*) of its own subsegments, but it is not a subsegment *in virtue of* its having subsegments. Rather, each is a subsegment *in virtue of* being a segment between the endpoints of the given segment that is not equivalent to the given segment. That explains why it is a subsegment. My point is that necessary conditions, even those that entail the existence of a constituent, are not necessarily part of explanatory or in-virtue-of conditions. In other words, "A holds only if B holds" can be true without "A holds in virtue of B holding" being true. (2003: 722)

Adopting Klein's terminology, we can say that R_1 being justified in the scenario above is not merely a necessary condition for P being justified but is an explanatory (in-virtue-of) condition for P being justified: it is a part of the explanation why P is justified. As I understand it, (IJC) specifies not only the necessary conditions but also the explanatory conditions for an inferential justification of a belief. What it says is to be understood along the following lines: if a given belief is inferentially justified, then it is inferentially justified *in virtue of* the fact that it is supported (at least) by another belief and the fact that that other belief is justified. The "if and only if" condition involved in (IJC) is to be conceived as an explanatory condition.

The upshot of this section is this. (RR) is the thesis that epistemic infinitism takes at face value what is suggested by the reason-giving game. (RR) is true simply because the reason-giving game suggests

that the justification of one and the same belief requires an infinity of reasons. And, what explains why the reason-giving game suggests this is (IJC) or, more particularly, the justification condition for inferential justification (viz. a belief is inferentially justified on the basis of another belief only if that other belief itself is justified, where the "only if" is meant to capture a sort of explanatory dependence).

6. *RS and the regress skeptic*

The question I now want to answer is whether (RS) is true. In this section, I will argue that even if it is true that the reason-giving game for a given belief between the regress skeptic and the subject satisfying the infinitist criteria ought rationally to result in a tie, skeptical attacks from the regress of justification are still not circumvented. If so, (RS) is also false.

Suppose that I and the regress skeptic have been playing the reason-giving game for my belief that P for quite a while, and we have left, say, thousands of steps behind, and both of us have started to lose patience. The skeptic recognizes that I have skillfully managed to cite an adequate reason for each belief that I have so far asserted and now openly concedes that I deserve a tie in the reason-giving game, that he cannot win the reason-giving game. This is a concession that the skeptic cannot succeed by continuing to raise the "what reason?" question in rationally concluding that my target belief is not justified. But now the skeptic realizes a crucial fact about the structure of my reasons, which he concedes to be infinite and non-repeating, and decides to change his strategy. Rather than continuing pointlessly to raise the "what reason?" question, the skeptic argues as follows:

> Look, I agree that you are in a position to provide an adequate answer to every "what reason?" question I raise for your beliefs. This is a remarkable feat; and to be honest, I was not expecting this. But now I realize that your victory is Pyrrhic, one that in effect signals your demise. You must agree with me that your belief that P can be justified on the basis of the reason you cite in its support only if that reason itself, which I grant you believe in, is justified, and this reason you cite in support of P is justified on the basis of another reason you cite in its support only if that latter reason itself, which I grant you believe in, is justified, and so on. This is just to take note of the fact that there is justification condition for inferential justification. Now, combine this with my concession that the structure of your reasons are infinite, and we get the result that your belief that P is not justified. This is because, given the infinity of the structure, the justification condition is never satisfied for your belief that P—the "only if" (as an explanatory dependence condition) is never eliminated or discharged: what we get is an infinity of conditionals structured

> like e_0 *is justified only if* e_1 *is,* e_1 *is justified only if* e_2 *is, and so on*, and it is clear that one can never get that e_0 is justified from such a structure.

Let me call this argument *the argument from the justification condition*. The argument is an old one, various versions of which have been presented by a number of philosophers in the past. To take just a few examples, the central point of the argument is made, sometimes metaphorically or cryptically, by the following remarks:

> The mode of reasoning based upon the regress ad infinitum is that whereby we assert the thing adduced as a proof of the matter needs a further proof, and this again another, and so on ad infinitum, so that the consequence is suspension, as we possess no starting point for our argument. (Sextus Empiricus 1976: 166)

> If there is a branch with no terminus, that means that no matter how far we extend the branch the last element is still a belief that is mediately justified if at all. Thus, as far as this structure goes, whenever we stop adding elements we still have not shown that the relevant necessary condition for the mediate justification of the original belief is satisfied. Thus the structure does not exhibit the original belief as mediately justified. (Alston 1986: 82)

> Consider a train of infinite length, in which each carriage moves because the one in front of it moves. Even supposing that fact is an adequate explanation for the movement of each carriage, one is tempted to say, in the absence of a locomotive, that one still has no explanation for the motion of the whole. And that metaphor might aptly be transferred to the case of justification in general. (Harkinson 1995: 189)

The argument from the justification condition purports to show that epistemic infinitism is not a non-skeptical alternative: if my beliefs are structured in the way the epistemic infinitist says they must, then none of those beliefs are justified because the justification condition for the inferential justification of each of those beliefs on the basis of (some of) the rest of my beliefs is *never* satisfied. Suppose that P is the proposition I believe whose justificatory status is in question. Suppose further that the justificatory status of P depends upon the justificatory status of R_1 (which I believe and provides evidential support for P), and the justificatory status of R_1 depends on R_2 (which I believe and provides evidential support for R_1), and so on to infinity in a non-repeating way. If this is so, then any member of this chain of propositions is justified only if the next member upon whose justification the justification of that member depends is justified. Since for each member in the chain there is another member upon whose justification its justification depends, and since there is no final member in the chain, none of the members of the chain is justified. The justification of each proposition in the chain involves a promissory note that is never paid, postponed to infinity. If so, (RS) is false.

The argument from the justification condition can be reformulated as a *reductio ad absurdum*. Suppose that P is justified. Where does its justification come from (or what does it depend on)? From (or on) R_1.

But where does R_1's justification come from? From R_2. So, we can say that P's justification comes from R_1 in the first instance and from R_2 in the second instance. Now where does R_2's justification come from? From R_3. So, P's justification comes from R_1 in the first instance and from R_2 in the second instance and from R_3 in the third instance. But now the question is: where does P's justification come from in the *last* instance? If P is justified, there must be an answer to this question: its justification must *ultimately* come from somewhere. This is because nothing can come from somewhere if it does not ultimately come from anywhere. Since given infinitism there is no answer to this question (there is no last instance), we arrive at the contradiction that P is both justified and unjustified. If this is so, the hypothesis that gives rise to the contradiction (i.e., *P is justified*) should be rejected.

7. *Infinitist responses considered*

My main aim here is not to argue that the argument from the justification condition is decisive but, more modestly, to argue that it is available to the regress skeptic willing to adopt the skeptical outcome approach but realizing that the reason-giving game played with a subject satisfying the infinitist criteria for justification ought to result in a tie. However, one might still reasonably wonder how strong the argument is or what the responses available to the infinitist are. In this section, I will address two objections the infinitist might level against the argument from the justification condition.

According to the first objection, the argument from the justification condition takes for granted a particular conception of inferential justification, one that the infinitist is not entitled to endorse. According to this conception, the structure of inferential justification is linear and the primary bearers of inferential justification are individual propositions (rather than systems of propositions). On this conception, a proposition's being inferentially justified is a property it might possibly have solely in virtue of another's proposition's transferring to it whatever justification it antecedently has thanks to there being suitable evidential relations between the two propositions. However, the infinitist might reject the linear conception of inferential justification and opt for accounting for the justification of a proposition on the basis of its relations to other propositions by adopting a holistic conception. In fact, this is what Klein qua the arch-infinitist exactly offers. Klein's infinitism is "warrant-emergentist" (2005a: 136), according to which justification is not property that can be *transferred* from one proposition to another but rather is a property that emerges whenever there is an endless, non-repeating sets of propositions available as reasons. Warrant-emergentist (or holistic) infinitism holds that "Being justified…is not a troublesome dependent property because a proposition being justified…does not arise in virtue of another proposition being justified—a proposition is justified for S in virtue of being a member

of a set of propositions each member having the required properties" (2003: 723).[14]

The skeptic might grant that the argument from the justification condition conceives the justification of a proposition on the basis of its relations to other propositions on the model of a transfer-account of justification, a model that takes justification-conferring relations to be linear rather than holistic; and, he might therefore grant that there might be some versions of infinitism, Klein's version being an example, that escape its threat. However, the skeptic might now wonder, quite plausibly I think, what motivation or rationale there is for adopting *holistic* infinitism. The crucial point is that what generates the regress of reasons *is* the linear conception of inferential justification: it is because a proposition, if it is inferentially justified, can receive its justification from another proposition that already has it that we are off to a regress of reasons each of whose justification depends on that of its successor. Infinitism is the view that fully embraces the regress of reasons that ensues from the linear conception of inferential justification. And, if the linear conception is abandoned, then it is unclear whether there is any good rationale for holding that justification requires an infinitely long sequence of reasons—the entire rationale that one might possibly have for preferring infinitism over its alternatives seems to be severely undermined. The moral is that *holistic* infinitism escapes the argument from the justification condition at the cost of undermining what might make *infinitism* an attractive option in the first place.[15]

According to the second objection, the argument from the justification condition rests on a failure to distinguish a *local* explanation of the justification of a particular proposition from a *global* explanation of why there are any justified propositions at all.[16] Suppose I want to explain why this billiard ball is moving now. Here what is to be explained is a particular event, the motion of this particular billiard ball. It seems that I can make reference to another particular event, Mr. Billiard's hitting that ball with his cue, in order to explain (at least partially) why it is moving. This is a local explanation of the motion of the ball, one that is purported to explain that particular event. However, suppose now that I want to explain why there is motion *at all*, why there is some motion rather than none at all. If this is the case, then I seek a global explanation of the very fact that there are things that move. The billiard ball is among the things that move; but if the explanandum is why there is motion at all, it seems that an appeal to Mr. Billiard's hit-

[14] Klein also writes: "The infinitist, like the coherentist, takes propositional justification to be what I called an emergent property that arises in sets of propositions. In particular, the infinitist holds that propositional justification arises in sets of propositions with an infinite and non-repeating structure such that each new member serves as a reason for the preceding one" (2007: 26).

[15] For further discussion, see, for instance, Demircioglu (2018).

[16] The distinction is widely discussed in the literature on the cosmological argument for the existence of God. I borrow it from Cameron (2018).

ting the ball is inadequate as an explanation of the motion of the ball since that hitting is an action that causes the motion of the billiard ball in virtue of the motion it itself exhibits. Such an "explanation" appears to be blatantly circular in its attempt to provide an explanans by an appeal to the explanandum itself. Armed with this distinction, the infinitist might now claim that the regress of reasons is purported to provide a local explanation of the justification of particular propositions but not a global explanation of why there are any justified propositions at all, and as such it is not threatened by the argument from the justification condition.[17]

There are a number of things the skeptic might say in response to this objection. First, the skeptic need not let the distinction between local and global explanations go unchallenged. The point of the distinction is to make room for local explanations of particular things of a certain kind (e.g., this moving ball) while admitting that there might be more to global explanations of the existence of things of that kind in general (e.g., things that are moving) than what local explanations can provide. However, it is not clear that there is really room for such a maneuver. It might be plausibly argued that one can only provide a local explanation of particular things of a certain kind if one can provide a global explanation of the existence of things of that kind. (Can I really explain the motion of this ball without being in a position to explain motion in general? Can the local explanation of the motion of that ball be divorced from the global explanation of motion in general?) Secondly, even if a distinction between local and global explanations can be plausibly drawn in the way suggested by the objection, it is not clear that a *philosophical* theory of knowledge and justified belief can rest satisfied with a local explanation of why a given particular belief is justified. A natural meta-epistemological view is that an epistemological theory aims to achieve a level of generality characterized by a sort of *global* worry: How can there be justified beliefs *at all*? If all infinitism has to offer is local explanations of why some particular beliefs are justified without a global explanation of how there can be justified beliefs at all, then it appears to be seriously incomplete as an epistemological theory of knowledge and justified belief.

I do not for a moment presume that this is the end of the dialectic between the regress skeptic and the infinitist, but I hold that the dialectic so far attests to the power of the skeptical outcome approach the regress skeptic might adopt by endorsing the argument from the justification condition.

[17] Klein (2003) suggests that this response is available to the infinitist but does not explicitly endorse it.

8. *Conclusion*

To sum up the discussion above, here then lies what I think is an ultimate tragedy of epistemic infinitism. Epistemic infinitism is suggested by the rules governing the reason-giving game as a proper response to the skeptic: the only way for us not to lose the reason-giving game against the skeptic is by our having at our disposal an infinity of reasons structured in a certain way. And, the reason why epistemic infinitism is suggested by the rules governing the reason-giving game is that there is justification condition for inferential justification: simply citing a reason in support of a belief is not enough to justify it, the subject must also be justified in believing the reason she cites. However, the justification condition for inferential justification can be deployed in an argument that epistemic infinitism fails to deliver a non-skeptical result. So, what makes epistemic infinitism come out as a viable option against the skeptic in the reason-giving game (namely, the justification condition) also renders it susceptible to a powerful skeptical assault. It is true that if our beliefs are structured in the way the epistemic infinitist says they must, then we do not lose the reason-giving game against the regress skeptic. But, despite this, I have argued that the skeptical outcome approach is still very much alive.

References

Adler, J. 2002. *Belief's Own Ethics*. Cambridge: MIT Press.

Aikin, S. 2005. "Who is afraid of epistemology's regress problem?" *Philosophical Studies* 126 (2): 191–217.

Aikin, S. 2008. "Meta-epistemology and the varieties of epistemic infinitism." *Synthese* 163: 175–185.

Aikin, S. 2009. "Don't fear the regress." *Think 8* (23): 55–61.

Aikin, S. 2011. *Epistemology and the Regress Problem*. New York: Routledge.

Alston, W. 1986. "Concepts of epistemic justification." In P. Moser (ed.). *Empirical Knowledge*. Towota: Rowman and Littlefield.

Brandom, R. 1994. *Making it Explicit*. Cambridge: Harvard University Press.

Cameron, R. 2018. "Infinite regress arguments." In E. Zalta (ed.). *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition).

Cling, A. 2004. "The trouble with infinitism." *Synthese* 138: 101–123.

Demircioglu, E. 2018. "Epistemic infinitism and the conditional character of inferential justification." *Synthese* 195 (5): 2313–2334.

Empiricus, S. 1976. *Outlines of Pyrrhonism*. Cambridge: Harvard University Press.

Fantl, J. 2003. "Modest infinitism." *Canadian Journal of Philosophy* 33 (4): 537–562.

Firth, R. 1978. "Are epistemic concepts reducible to ethical concepts?" In A. Goldman and J. Kim (eds.). *Values and Morals*. Dordrecht: Reidel Publishing Company.

Harkinson, R. J. 1995. *The Skeptics*. Routledge: New York.

Klein, P. 1998. "Foundationalism and the infinite regress of reasons." *Philosophy and Phenomenological Research* 58 (4): 919–925.

Klein, P. 1999. "Human knowledge and the infinite regress of reasons." *Philosophical Perspectives* 13: 297–325.

Klein, P. 2000. "Why not infinitism?" *The Proceedings of the Twentieth World Congress of Philosophy* 5: 199–208.

Klein, P. 2003. "When infinite regresses are not vicious." *Philosophy and Phenomenological Research* 66 (3): 718–729.

Klein, P. 2005a. "Infinitism is the solution to the regress problem." In M. Steup and E. Sosa (eds.). *Contemporary Debates in Epistemology*. Oxford: Blackwell Publishing.

Klein, P. 2005b. "Reply to Ginet." In M. Steup and E. Sosa (eds.). *Contemporary Debates in Epistemology*. Oxford: Blackwell Publishing.

Klein, P. 2007a. "Human knowledge and the infinite progress of reasoning." *Philosophical Studies* 134: 1–17.

Klein, P. 2007b. "How to be an infinitist about doxastic justification." *Philosophical Studies* 134: 25–29.

Klein, P. 2011. "Infinitism." In S. Bernecker and D. Pritchard (eds.). *Routledge Companion to Epistemology*. New York: Routledge.

Klein, P. 2014. "No final end in sight." In R. Neta (ed.). *Current Controversies in Epistemology*. New York: Routledge.

Leite, A. 2005. "A localist solution to the regress of epistemic justification." *Australasian Journal of Philosophy* 83 (3): 395–421.

Rescorla, M. 2009. "Epistemic and dialectical regress." *Australasian Journal of Philosophy* 87 (1): 4–60.

Turri, J. 2010. "On the relationship between propositional and doxastic justification." *Philosophy and Phenomenological Research* 80 (2): 312–326.

Wright, S. 2013. "Does Klein's infinitism offer a response to Agrippa's trilemma?" *Synthese* 190: 1113–1130.

Croatian Journal of Philosophy
Vol. XX, No. 58, 2020

Are people smarter than machines?

PHIL MAGUIRE, PHILIPPE MOSER and REBECCA MAGUIRE
National University of Ireland, Maynooth, Ireland

Recent progress in artificial intelligence has led some to speculate that machine intelligence may soon match or surpass human intelligence. We argue that this understanding of intelligence is flawed. While physical machines are designed by humans to simulate human rule-following behaviour, the issue of whether human abilities can be emulated is not well-defined. We outline a series of obstacles that stand in the way of formalizing emulation, and show that even a simple, well-defined function cannot be decided in practice. In light of this, we suggest that the debate on intelligence should be shifted from emulation to simulation, addressing, for example, how useful machines can be at particular tasks, rather than deliberating over the nebulous concept of general intelligence.

Keywords: Artificial intelligence, Turing test, Church-Turing thesis, technological singularity, simulation, Turing machines.

1. *Introduction*

Could a human-made machine ever surprise its creator, by taking initiatives of its own? According to Cristianini (2016), this is a question that has been asked for centuries, resulting in a variety of answers. Arguably the first computer programmer, Ada Lovelace knew where she stood on this issue: "The Analytical Engine has no pretensions whatever to originate anything", she stated in 1843. "It can follow analysis; but it has no power of anticipating any analytical relations or truths".

And yet, 173 years later, a computer program developed just over a mile from her house in London beat Lee Seedol, a 9-dan master of the game Go. None of AlphaGo's programmers could ever hope to defeat Lee Seedol, let alone defeating their own program. According to Cristianini (2016), the software has learned to do things its programmers can't do and don't understand. The machine learning techniques used by AlphaGo are becoming widespread in the field of AI. Whereas in the past the idea of a "learning machine" might have sounded like a con-

tradiction, it now seems reasonable to speak of physical machines that are flexible, adaptive, or even curious (Cristianini 2016).

Given these recent breakthroughs, some commentators have suggested that machine learning will continue to improve to the point where it surpasses human ability in many domains. At the Future of Humanity Institute's conference on machine intelligence in 2011, Sandberg and Bostrom (2011) conducted an informal poll eliciting the views of participants. The median estimate for the emergence of human-level machine intelligence was the year 2050 (see also Müller and Bostrom 2016).

We argue in this paper that the idea of AI somehow surpassing humanity constitutes a misrepresentation of the nature of intelligence. The goal of machine learning is not to recreate intelligence or even to define it, but instead to show that many tasks that were previously assumed to require human intervention can be successfully automated. The pertinent question is not whether machines are going to overthrow humanity in a technological singularity (e.g. Bostrom 2014), but how resistant different aspects of human behaviour are to simulation.

In order to make this case we highlight a series of obstacles that lie in the way of formally defining the concept of emulating human intelligence. We then show that even simple, well-defined functions cannot be decided in practice. Returning to the question of whether people are smarter than machines we conclude that, rather than grappling with the concept of general intelligence, philosophers should instead focus on anticipating the utility that machines might provide on particular constrained tasks.

2. *Problems with the question of emulation*

Building on earlier work by Kurt Gödel, the theory of computation was independently discovered from different perspectives in 1936 by Alonzo Church and Alan Turing. Church's version was based on the λ-calculus, while Turing's was based on what is now known as the Turing Machine. In his 1936 article, Turing presents the idea of a Universal Turing Machine (UTM), which is capable of simulating any other Turing Machine. The key ingredients for this breakthrough are:

1) the idea of capturing general recursive functions (a.k.a. computable functions) in the form of a simple model for symbol manipulation (a.k.a. Turing Machines)
2) the philosophical position that general recursion captures all effective methods, a position now known as the Church-Turing thesis

With these two ingredients, all effective processes can be enumerated. This enumerability supports the concept of universal computation, as it allows a single "Universal Turing Machine" (UTM) to read in a description of any other effective method to be simulated, as represented

by the machine's index in the ordering of Turing Machines. A UTM is thus capable of simulating any process that is effectively calculable.

This result seems to open the door to human-level AI: a single physical machine can be developed which can, when given the appropriate program, simulate every effective process conceivable. Could a UTM running a specially-designed program therefore emulate the 'program' running in the human brain? We identify several obstacles that lie in the way of formalizing and then deciding such a question.

2.1 *We don't know what it means to build a Turing machine in practice*

The UTM is an abstract mathematical idea. Physical machines are engineered to offer only finite precision, rather than the potentially unbounded precision and memory space required by Turing's definition. Thus, physical machines merely *simulate* the behaviour of a genuine Turing machine. This issue goes beyond the lack of an infinite tape, it concerns the very mechanics of the device: we simply can't be sure that a physical machine will continue to compute properly without making mistakes at some point in the future due to unforeseen engineering limitations. In the words of Wittgenstein (RPP I 1096), Turing's 'machines' are, actually, "humans who calculate". Turing (1948/9) himself clarifies that human behaviour sets the standard for his concept: "A man provided with paper, pencil and rubber, and subject to strict discipline, is in effect a universal machine."

Physical machines are engineered by humans to *simulate* human computing abilities with a certain level of fidelity. They are not intended to emulate the standard of computation benchmarked by humans. By 'emulate' we mean to match or exceed human capacity, as opposed to 'simulate', which involves a finite level of success at imitation:

> The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer. (Turing 1950)

Also:

> Electronic computers are intended to carry out any definite rule of thumb process which could have been done by a human operator working in a disciplined but unintelligent manner. (Turing 1950b)

The dual interpretation of the word 'machine', as in Turing machine (human abstraction) versus computing machine (physical device), can lead to confusion. The 'machine' in the term 'Turing machine' refers to the idea of strict rule following without imagination. It does not refer to a physical device. Humans compute in a way which is captured by the abstract idea of a Turing machine, whereas electronic devices are engineered to simulate that behaviour. Human-built machines represent only a finite amount of engineering calibration carried out by a relatively small set of humans. Whereas human consensus sets the standard for computation, electronic devices merely simulate that abil-

ity to a finite degree of precision, one which is continually improved by developments in technology (see Maguire and Maguire 2018).

Because Turing machines are a mathematical ideal, it remains unclear how such a concept could be represented by a physical object in practice.

2.2 *We can't formalize what it means for a UTM to be universal*

The cornerstone of computation is the concept of the stored program, the idea that no function is special, that every function can be represented as data inputted to a single UTM. However, this intuitive foundation depends crucially on the Church-Turing thesis (see Copeland 2002). The Church-Turing thesis asserts that a function on the natural numbers is computable by a human following an algorithm (ignoring resource limitations) if and only if it is computable by a Turing machine. In other words, it states that the idea of a Turing machine captures everything there is to the notion of a human following the step-by-step instructions of an algorithm.

The physical form of the Church-Turing thesis in even more restrictive. It states that a Turing machine captures every act of algorithmic rule-following that a human can achieve even when exploiting the use of exotic physical processes, such as black holes or quantum systems (see Cuffaro and Fletcher 2018; Earman 1993; Kieu 2004). Thus, the physical Church-Turing Thesis could be false without humans necessarily being able to demonstrate a superior ability using only pen and paper.

The concept of universal computation (i.e. the idea that a UTM exhausts the set of effectively computable functions) depends on the Church-Turing thesis. If the Church-Turing thesis turns out to be false, then Turing machines would lose their privileged position: there would exist logical processes that could not be simulated by any given Turing machine, but which could be computed by a human using some other rule-following process. There might not be any automaton capable of computing all the new functions, in which case the concept of universality would be lost.

As soon as we accept that an automaton is capable of universal computation, or that it supports a Turing-complete language, we are relying on the Church-Turing thesis. Nevertheless, we have no *proof* that the Church-Turing thesis is true, and, according to Turing (1954), have no aspirations of ever discovering such a proof (cf. Black 2000; Dershowtiz and Gurevich 2008). By definition, the thesis seems to be outside the scope of proof, because it speaks about the set of effective methods, and the process of proof-checking is in that set. Although there are no known counter-examples, and different formalisms converge towards the same result, the Church-Turing thesis is not the type of statement we aim to prove. It exists as an informal statement in natural language, not in a form that could be processed by a Turing machine. So, is the thesis 'true'?

Certainly, there has been no refinement in the notion of effective method since 1938, when Kleene refined Turing and Church's (1936) method by applying it to partial functions. In that sense, there is convincing evidence that it is *hard* to identify a stronger notion of effective method than that provided by Church and Turing. And yet, it is not possible to guarantee that, at some point in the future, a new development in the understanding of human rule-following abilities will reveal a limitation in the Turing machine's functionality that vitiates its putative property of exhausting the set of effective methods. We simply don't know. Although the concept of universal computation is intuitively convincing, its existence is not something that has been formally proved.

Although Church and Gödel were happy to accept the Church-Turing account of effective method as a *definition* (according to Gödel "... the correct definition of mechanical computability was established beyond any doubt by Turing"), Emil Post, who in 1936 delivered an alternative model of computation, was vociferous in his opposition. According to him, "to mask this identification under a definition hides the fact that a fundamental discovery in the limitations of mathematicizing power of Homo Sapiens has been made and blinds us to the need of its continual verification". Post hoped to publish a series of "wider and wider formulations ... The success of the program would for us, change this hypothesis not so much to a definition or to an axiom but to a natural law" (Post 1936).

Turing adopted the middle ground, accepting computation's strong intuitive foundation, while at the same time acknowledging that the thesis would always remain unproved. In 1954 he remarked: "The statement is...one which one does not attempt to prove. Propaganda is more appropriate to it than proof, for its status is something between a theorem and a definition."

In sum, the Church-Turing thesis remains an informal statement, not a mathematical one. It cannot be fully formalized, and consequently it cannot be processed by a computing machine. It is a sophisticated thesis expressed in natural language, on which the universality of a UTM depends, yet it lies beyond the scope of a UTM.

This presents another obstacle to the emulation of human intelligence. The recognition of emulation depends on the recognition of universality, which itself hinges on the truth of a sophisticated thesis in natural language, which cannot be formally addressed by a Turing machine.

2.3 *We can't formally address the potential existence of human abilities beyond computation*

Even if we could somehow formally confirm the Church-Turing thesis, there might still be some human abilities which remain beyond rule-based computation. In other words, a UTM might have some limita-

tions that humans are able to 'appreciate' in some nebulous sense that is itself beyond automation. Any claim that machines can emulate human abilities has to deal with this possibility.

For example, in his 1936 paper, Turing identified the existence of a well-defined problem which is beyond computation (i.e. an uncomputable problem). This allowed him to resolve in the negative the Entscheidungsproblem, the question of whether there exists an algorithm to decide whether a given statement in first order-logic is provable or not. Turing showed by contradiction that no such algorithm is possible. He imagined taking an automaton which decides if an algorithm will halt or not, and then feeding it a description of itself (now made possible by the enumerability of effective methods). This creative leap allowed him to precisely define an object, known as the halting language, which cannot be produced by any effective method whatsoever.

The issue here is that no Turing machine can ever 'know' that the halting language exists. To do that it would have to be able to appreciate that there is something it cannot do, without trying to do it. Thus, it *seems* that humans hold a privileged perspective over Turing machines, insofar as we 'know' that the halting language exists. For example, Lucas (1961) argued that humans can look and see that a given machine's Gödel sentence is true, meaning they can always do something that the machine cannot. This argument has been further developed by Penrose (1994) to show by contradiction that human abilities could never be formalized to the point at which a Gödel sentence becomes discernible.

Similarly, even though the halting language cannot be constructed or represented in nature, it seems to be defined for the human reader in a finite number of symbols by Turing's 1936 article. If humans can indeed appreciate such a definition, then they are capable of recognizing the idea of an object that no computer can ever represent.

In his 1938 PhD thesis, carried out under the supervision of Church, Turing makes clear his view that the human mind has an intuitive power for performing uncomputable steps beyond the scope of a Turing machine:

> Mathematical reasoning may be regarded rather schematically as the exercise of a combination of two faculties, which we may call intuition and ingenuity. The activity of the intuition consists in making spontaneous judgments which are not the result of conscious trains of reasoning...In consequence of the impossibility of finding a formal logic which wholly eliminates the necessity of using intuition, we naturally turn to non-constructive systems of logic with which not all the steps in a proof are mechanical, some being intuitive.

After the Second World War, Turing's view on the role of intuition in reasoning appears unchanged. In a 1948 report to the National Physical Laboratory, Turing again clarifies that mathematicians' ability to decide the truth of certain theorems appears to transcend the methods available to any Turing machine:

> Recently the theorem of Gödel and related results...have shown that if one tries to use machines for such purposes as determining the truth or falsity of mathematical theorems and one is not willing to tolerate an occasional wrong result, then any given machine will in some cases be unable to give an answer at all. On the other hand the human intelligence seems to be able to find methods of ever-increasing power for dealing with such problems, 'transcending' the methods available to machines.

In his last article published before his death in 1954, Turing again emphasises the role of intuition beyond effective method. He argues that Gödel's theorem shows that 'common sense' is needed in interpreting axioms, something a Turing machine can never demonstrate:

> The results which have been described in this article are mainly of a negative character, setting certain bounds to what we can hope to achieve purely by reasoning. These and some other results of mathematical logic may be regarded as going some way towards a demonstration, within mathematics itself, of the inadequacy of 'reason' unsupported by common sense.

Human intuition is an ability that resists formalization or description. It is not possible to verify the emulation of human abilities if we cannot even represent what those abilities are.

2.4 *Even simple, well-defined functions are undecidable in practice*

In sum, merely formalizing the question of whether physical machines can emulate human intelligence is fraught with great difficulty. And yet, even if it could somehow be formalized, the question would still not be a useful one, because it couldn't be answered in practice.

Let's assume that it's somehow possible to build a true physical Turing machine. Let's assume that we somehow 'know' for sure that Turing machines are capable of universal computation, and that all aspects of human behaviour can be described in terms of that computation. Even with all of these assumptions, it can be shown, using an argument from theoretical computer science, that the question of emulating a given program remains *undecidable*, since, in practice, even the simplest functions are undecidable from their output.

Below, we provide a modification of the use theorem (see Odifreddi 1992) to show that no finite set of interactions is sufficient for deciding what process, whether computable or uncomputable, is behind the behaviour of a black-box system: properties of functions cannot be decided in practice based on their output. No matter how many questions are asked, it is not possible to know for sure what is behind the output of a black-box system. The argument is related to that of Gold (1967), who showed that any formal language that has hierarchical structure capable of infinite recursion is unlearnable from positive evidence alone. It is also related to Rice's theorem (1953), which shows that all non-trivial semantic properties of programs are undecidable (in the case of Rice's theorem access is given to the program code itself, rather than its output).

More formally, let O be an observer, A a set of strings over some finite alphabet Σ, and f: $\Sigma^* \rightarrow \{0,1\}$, our black-box, be a Boolean function.

O can adaptively ask finitely many queries $f(x)$=? (O has access to A), after which O decides whether f computes the set A, i.e. $f(x)=A(x)$ for every x.

The following standard argument shows every observer is wrong on some function (i.e. the past behaviour of the black-box cannot be used to decide its future behaviour).

For any observer O, and any set A, there is a function f: $\Sigma^* \rightarrow \{0,1\}$ *such that O is wrong on f.*

Proof.
Let O be as above, A be a set. If O rejects all functions (i.e. thinks all functions do not compute A) then O is wrong on f, where $f(x)=A(x)$ for every x. So let g be accepted by O. O queries g on finitely many strings $x_1,x_2,\ldots, x_n$. On all the strings $x_1,x_2,\ldots, x_n$, g is equal to A, otherwise O is wrong about g. Choose y different from $x_1,x_2,\ldots, x_n$, and construct f: $\Sigma^* \rightarrow \{0,1\}$, by letting $g(x)=f(x)$ for all $x \neq y$, and $f(y)=1\text{-}A(y)$. f does not compute A, because f is different from A on input y. Because f equals g on inputs $x_1,x_2,\ldots, x_n$ (the ones queried by O), O will make the same decision about f as about g, i.e. O decides that f can compute A. By construction of f, O is wrong.

In sum, this result shows that not only is a finite interaction incapable of deciding intelligence, it is not even capable of deciding any function whatsoever. Even an oracle with access to the halting language could not produce behaviour which reliably separates it from a simple Turing machine. Past behaviour is never sufficient for deciding whether a system is doing something smart. A computable process can mimic an uncomputable one up to any finite duration of interaction.

The question of emulating human intelligence thus holds no utility in practice. No finite set of interactions can always be relied on to expose the lack of intelligence of any given machine. If a black-box system has not yet made a mistake, there is no way to tell whether or not it will make any mistakes in the future. Thus, the propensity to make mistakes at some stage *does not matter*.

According to Turing (1948): "the condition that the machine must not make mistakes ... is not a requirement for intelligence". At the infinite limit, mistakes are inevitable, but in practice those mistakes can be pushed back as far as one wants. Turing (1947), in his earliest surviving remarks concerning AI, points out that this would allow machines to play very good chess:

> This...raises the question 'Can a machine play chess?' It could fairly easily be made to play a rather bad game. It would be bad because chess requires intelligence. We stated... that the machine should be treated as entirely without intelligence. There are indications however that it is possible to make the machine display intelligence at the risk of its making occasional serious mistakes. By following up this aspect the machine could probably be made to play very good chess.

Rather than dismissing the idea that humans are ultimately smarter than machines, Turing (1950) is instead highlighting the lack of practical significance of such an idea: in the physical world there will always be some machine which is up to the job of simulating intelligence to a required finite length before making any mistakes:

> There would be no question of triumphing simultaneously over all machines. In short, then, there might be men cleverer than any given machine, but then again there might be other machines cleverer again, and so on.

It should be noted that reducing human behaviour to the computation of a function is already a very strong modelling assumption, even before the issue of emulation is tackled. The observable behaviour of some physical systems, such as chaotic deterministic systems, cannot be described by the computation of any function (see Longo and Paul 2011). Applying a functional description to biological individuals would no doubt be close to impossible.

In practice then, emulation is a thoroughly useless idea, or in Turing's (1950) words, an idea "too meaningless to deserve discussion". Attributing intelligence to any being or object with certainty is an undecidable issue at best. Whatever intelligence is, it's not something that depends on confirmation. So what is it? In the remainder of the paper we examine alternatives to emulation.

3. *Emulation is not a useful concept, but simulation is*

Thus far we have highlighted why the question of a machine emulating human abilities has no utility. But if the emulation of intelligence holds no utility, does this imply that intelligence is itself a useless idea? What, if anything, does the concept of intelligence imply in practice? One possible conclusion is that intelligence has no discernible real-world effects whatsoever, having nothing to do with behaviour.

This is the attitude adopted by Professor Jefferson in his 1949 Lister Oration (which Turing was responding to in his 1950 article): "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain—that is, not only write it but know that it had written it."

Here, Jefferson is arguing that behaviour alone is never sufficient for providing evidence of intelligence. Instead, we must 'know' what words mean and 'feel' emotions. Because such properties can never be represented symbolically, there is no possibility of any system, human or otherwise, evidencing its intelligence in practice. Intelligence and behaviour have no relationship at all.

But this doesn't seem right. Intuitively, what we mean by 'intelligence' is something useful. Our human abilities let us achieve things in the real world that simple rule-following systems could not. It seems as though we can quickly and reliably identify intelligence through the communication of symbols. For example, this article is only a few pages

long, yet (we hope) it strongly suggests an intelligent origin. It seems feasible that a finite signal beamed from a distant solar system could convince us that it harbours intelligent life. We could never be absolutely 100% sure, but it seems plausible that there exist signals that could lead us to be very, very confident.

Indeed, all the communication that has ever taken place between human beings can be summarized as a finite string of symbols. Human communication, of course, relies not just words, but also gesture, voice tone, facial expression, body language and context. Assuming a continuous high fidelity recording of what an individual sees and hears, all of this information could be translated into a finite set of 1s and 0s. If intelligence could not be evidenced in practice through finite interactions, it would preclude humans from identifying each other as intelligent, reducing us to solipsists.

It seems that in order for the concept of intelligence to be a meaningful and useful one, there must be some practical means of identifying and engaging with intelligent systems in the real world. Having realised this, Turing (1950) remarks "I am sure that Professor Jefferson does not wish to adopt the extreme and solipsist point of view. Probably he would be quite willing to accept the imitation game as a test."

Intelligence does something in practice. Although it cannot be used to *decide* the intelligent origin of a signal (i.e. choose yes or no with absolute certainty), it seems as though it can be used to detect the footprint of intelligence with high confidence. According to Aaronson (2006), "people regularly *do* decide that other people have minds after interacting with them for just a few minutes...there *must* be a relatively small integer n such that by exchanging at most n bits, you can be reasonably sure that someone has a mind" (see also Harnad 1992).

With this in mind, Turing (1950) makes clear that the interesting question is not whether machines can emulate humans (an undecidable proposition at very best), but *how difficult* it will be to build useful machines that simulate human behaviour closely, for extended periods of time. Specifically, the questions about intelligence that can be meaningfully asked and answered are those concerning how resistant different human abilities are to simulation.

Turing switches the focus from emulation to the simulation of intelligent behaviour, describing the idea of an imitation game, a 'test' which sets human behaviour as the standard to be simulated for a finite duration. The goal is for machines to confound the heuristics that people typically rely on for detecting signals of intelligent origin. Hodges (2009) succinctly expresses this idea: "operations which are in fact the workings of predictable Turing machines could nevertheless appear to the human observer as having the characteristics of genuine intelligence and creativity".

To be clear, Turing's test is not a test for *deciding* intelligence. Turing (1950) never once refers to machines 'passing' his test; the test is not intended to provide evidence of anything beyond the ability of a machine to do well at that test for a finite period of time, hence the notion of 'passing' doesn't hold any particular significance. The imitation game merely provides a vehicle for quantifying how resistant human behaviour is to simulation (albeit, an unreliable one). If a machine passes one test, we do not deduce anything further about the abilities of that machine, because there is no guarantee whatsoever that it will pass another test. Instead, we conclude that the test is not as hard as we thought it was, that it is perhaps no longer a strong test for intelligence. The test is not intended to address the question of whether machines can emulate intelligence (i.e. that they could simulate human behaviour perfectly for any length of time). Instead, Turing's (1950) contribution is to take a question "too meaningless to deserve discussion" (i.e. "Can machines think?" / can machines emulate human abilities?) and to transform it into a meaningful question that can be addressed in practice (how resistant are human abilities to simulation?).

Turing seeks merely to establish the possibility of "satisfactory performance" at the imitation game over a finite period (i.e. finite simulation); not perfect performance, nor the idea that satisfactory performance is a perfect predictor of subsequent perfect performance. He never goes beyond claims for the finite simulation of intelligence: "My contention is that machines can be constructed which will simulate the behaviour of the human mind very closely" (Turing 1951).

4. *Designing a good test*

Some have interpreted Turing (1950) as suggesting that infinite testing is required to establish intelligence, spread over an infinite length of time (e.g. Harnad 1992). Again, Turing's focus is not on establishing that machines can emulate human thinking, a concept which he describes as "meaningless". Instead, he is speculating on the difficulty of identifying reliable tests for discriminating human intelligence. Even if humans have intuitive abilities beyond machines, it may be difficult to demonstrate such abilities *in practice*. How hard is it to identify a reliable test for intelligence?

Let's consider the question of "what is a good test"? A test is of finite length. Applying it to an object yields results that enable inferences to be drawn about that object. Somehow, the results hold significance for other aspects of the object, beyond those which have been directly tested. One could say that the test succeeds in succinctly 'characterising' the object through a finite set of investigative results.

For example, students are asked to sit tests to reveal how much they know about a particular subject. Because of the short duration, it is not possible to ask them every question that could possibly be asked. Instead, questions are chosen cleverly so that responses can be relied

on to draw inferences about students' ability to answer all the other potential questions which haven't been asked. A good test allows the tester to make inferences about future behaviour based on past behaviour.

Of course, a particular student might get lucky on a test. They might fortuitously (or by cheating) have learned off the answers to the exact questions which came up, but no others. Thus, as previously argued, a test can never *decide* whether a student fully understands a subject. What a cleverly crafted test can do is offer a very high level of confidence that the student would have answered other questions correctly. Past behaviour can allow us to predict future behaviour with high confidence.

What are the properties of a good test that would lead us to have such confidence? In short, a good test is one for which there is no easy strategy for passing it, other than full mastery of the subject. For a start, there should be no way for students to get a copy of the test in advance, or predict what will be on it so that they can learn off the relevant responses without understanding the subject deeply. In addition, the test should be well diversified, bringing together material from many different areas of the subject. For instance, the answers should draw on different aspects of understanding, and not betray a simple pattern which would allow them to be all derived using the same technique. Finally, successive answers should be integrated with each other, rather than addressing separate chunks of knowledge which could be learned off independently. When one answer builds on the next, the only way to do well is to understand everything.

These criteria for test reliability can be summarized as follows: the content of the test should be *random* relative to the set of questions that could potentially be asked, and also internally integrated, so that questions cannot be answered independently of each other. If we follow these criteria, it seems the difficulty of passing the test can increase exponentially relative to its length.

If test questions were leaked in advance, then machines would only need to hardcode the appropriate responses to ensure success. How can we ensure that test questions are as unpredictable as possible? Turing's (1950) idea is to hand this responsibility over to human judges. Given that they can rely directly on their own intelligence (whatever that is), questions derived by humans on the fly have the potential to be hard to answer. In addition, human judges have the greatest ability to integrate subsequent questions into the preceding conversation, so as to ensure there is no trivial algorithm for computing the relationship from input to output.

Of course, this only applies in the best case scenario, when human judges choose the most challenging questions conceivable. But how do we know which questions are reliable indicators of intelligence?

Although intelligence might give us the ability to pass convincing tests easily, it does not necessarily give us the ability to easily find, generate or recognize good tests. Given a particular test, how can we *know*

that there is no simple program that quickly computes the answer? For example, the Winograd schema challenge (e.g. Levesque 2014) currently poses great difficulty for machines, yet we have no guarantees that developments in AI over the next few years will render such problems obsolete.

Whenever researchers put forward what intuitively appears to be a challenging test for AI, such as playing chess, nobody knows for sure how hard it really is. Problems that are assumed to require deep insight can end up having relatively simple mechanical solutions. For example, Hofstadter (1980) erroneously predicted that no dedicated program would ever defeat a chess champion, because playing the game well constituted a test for general intelligence:

> "Do you want to play chess?" "No, I'm bored with chess. Let's talk about poetry". That may be the kind of dialogue you could have with a program that could beat everyone. (Hofstadter 1980)

Turing's concept of a test is not intended as a once-off decider of emulation. Instead, it represents the idea of a never-ending battle to establish the superiority of human intelligence over rule-following. Turing believed that it would be a battle in perpetual retreat, with supposedly reliable tests continuing to fail:

> It is customary, in a talk or article on this subject, to offer a grain of comfort, in the form of a statement that some particularly human characteristic could never be imitated by a machine...I cannot offer any such comfort, for I believe that no such bounds can be set. (Turing 1951)

Inevitably, if a Turing-style test is run using laypeople, the programs that get furthest will be those that exploit the weaknesses of human psychology. People who aren't trained in AI can be more easily fooled. Thus, rather than being inferred from the length of questioning, resistance to simulation could be quantified by unrestricted open competition, with significant prize money awarded to expert machine-exposing teams. Turing seemed to assume that the tester would be at the level of an informed graduate of Oxford or Cambridge (McDermott 2015). Either way, Turing's opinion was that coming up with new tests that reliably separate people from machines was going to get harder quite quickly.

Once a test is found to be passed by a machine, then the test is busted. It can no longer be relied on to provide evidence of intelligence. As soon as a machine succeeds in defeating a test, researchers go back to the drawing board to develop a harder test. The process never ends. In the same way that it is not possible to decide intelligence, it is not possible to decide the reliability of a test for intelligence. Tests must themselves be tested, in an unending cycle of doubt.

Even though the question of emulating human abilities is useless, the practical issue of evidencing an ability gap between machines and humans is becoming more and more challenging. This explains why Turing (1950) was upbeat on the imminent prospect of artificial intel-

ligence. The behaviours that were intuitively assumed to be reliable tests in 1950, such as playing good chess, or engaging in convincing conversation, had never been exposed to machine-scrutiny before, making Turing quite confident that they would not hold up for long. This confidence is evident in his prediction that by the end of the 20th century people would "be able to speak of machines thinking without expecting to be contradicted".

5. *Are people smarter than machines?*

Who is better at the game of Go, humans or physical machines? Although there will probably never again be an individual human that can defeat or improve on the top Go-playing program, we expect that humanity as a whole will continue to overthrow every reigning program by constantly designing better and better ones.

For instance, humans have already created an improved computer program that is capable of beating AlphaGo, called AlphaGo Zero. Developed within 2 years of its predecessor, it beat the original version of the program 100 games to 0 (see Silver et al. 2017). If humanity as a whole retains the ability to improve on the world's top Go-playing software, then humanity must know something about the game of Go that the software does not. While the individual AI developers who together created AlphaGo Zero are individually beaten by their collective creation, they can, when working together, find ways to improve on the state of the art. Technology and AI offer a way for humans to combine and leverage their collective engineering prowess into a single system whose efforts can be focused on a specific problem which is beyond the understanding of any single person.

The main advantage that a program has over an individual human is being able to concentrate the historical wisdom provided by many different people over a lengthy period of time, and to apply it quickly. Machines can be faster, cheaper, more reliable, more durable, and can hold greater memory. However, this 'brute force' does not equate to emulation. Brute force can surpass human ability over short runs, but in the longer run humans have the ability to innovate superior algorithms for performing the same task even more quickly.

For example, just because current chess programs can beat any grandmaster in the world at chess does not mean that computers are better at chess than humanity as a whole. We can say that they simulate the intelligent playing of chess well, under certain conditions, for a certain period of time. But we still cannot show that a program *emulates* human ability at chess.

The output of any human-built machine simply reflects the stored historic work of humans, which always involves a *finite* amount of effort drawn from a potentially unbounded set. The possibility always remains for humans to carry out even more work, and build an even better machine, which more closely simulates human intelligence.

While highly engineered machines can 'simulate' human ability to avoid mistakes for a long time, such performance is never sufficient to rule out the possibility of some bug that was too rare to be anticipated. For this reason, at no point in the future will humans recognize the behaviour of a human-built machine as the ultimate authority for what counts as logically correct. Human-built machines will always have human errors embedded in their makeup that their original builders missed, but that the hardware and software engineers of tomorrow can fix. Consequently, humanity does not learn about logic by observing the activity of human-built machines.

Granted, an individual human can make a mistake relative to the standard held by a larger group of humans, or temporarily to a machine, but the whole of humanity cannot make a mistake relative to a machine. While physical machines may provide useful information to one person in a particular context, they never provide information to humanity as a whole: human-built machines merely represent a store of humanity's logical and engineering effort, which can then be reused and applied to novel problems.

From this perspective, we can see that the idea of physical machines surpassing humanity is nothing more than a clever trick. Physical machines can certainly impress individual humans, but only by recycling and cleverly blending the stored historical wisdom of larger groups of humans.

And yet ... any claims that humanity might make to ultimate superiority over machines reflect nothing more than useless intuitions. As noted by Turing (1950), assertions of the mind's superiority are "without any sort of proof". Intuitively we seem to 'know' that people are smarter than machines, but in practice that means nothing.

6. *Identifying useful questions for AI*

Interpretations of Turing's (1950) work have focused strongly on the idea of a specific challenge that, once passed, has significant implications. For example, Warwick and Shah (2015) claim that the Eugene Goostman chatbot machine "became the first to pass the Turing Test, as set out by Alan Turing, on unrestricted conversation". Turing's article has often been interpreted either as being supportive of functionalism (e.g. Searle 1980), or of advocating a trite, deeply flawed test for evaluating the intelligence of artificial systems through the process of imitation (e.g. French 2012). Shieber (1994), for instance, interprets Turing (1950) as making the claim that "any agent that can be mistaken by virtue of its conservational behaviour [for] a human must be intelligent" (see Copeland 2003, for further examples). Hayes and Ford (1995) go so far as to interpret Turing as proposing "a test of making a mechanical transvestite" and state that "Turing's vision from 1950 is now actively harmful to our field".

Although the specific test described by Turing is no doubt dated, we have argued that his article is not focused on emulation. Because humans cannot even express what it would mean for a physical machine to emulate human abilities, the question is useless. Instead, Turing was speculating about what physical machines of the future would be able to accomplish in practice. He hinted at, not a specific challenge, but a general thought experiment, involving the hypothetical *simulation* of human intelligence by imaginable computers. Although it currently seems as though we can quickly and reliably identify human intelligence through the communication of symbols, the same methods of discrimination we rely on now might not hold up in the future. Reliable tests may prove harder and harder to find. For instance, realistic sounding chatbots may end up phoning people and holding functional conversations without being identified as automata. Future software may be able to generate images, audio, and videos of humans that look and sound like real humans but are actually fake, indistinguishable from "real" digital representations of people. Even though it seems that humans possess some intuition beyond formal logic, it might still become quite difficult to separate humans from machines in practice.

Can we design a test for intelligence that can be run in practice, though not passed by a machine? Intuitively it seems like we should be able to. But the missing ingredient here is proof. While it seems like we can set and pass tests which reliably draw a line between us and machines, we cannot prove it. We cannot say anything definitive at all about the relationship between computable functions and intelligence beyond the realm of the computable. We cannot bridge that gap in any meaningful way. There's nothing useful that can be said about intelligence beyond seeking to simulate it bit by bit in practice, by continually improving our machines and seeing what they are capable of.

This basic insight allows us to separate questions about machine intelligence that are useful from those that are not.

How resistant is language translation to simulation?—We can ask this question. Specifically, we can use the imitation game to quantify the difficulty of developing a machine that simulates human-level language translation to some level of accuracy. The potential availability of an answer renders the question meaningful.

Is human-level language translation beyond machines?—This question has no utility because it is not well-defined.

Can machines do language translation better than humans?—This question has no utility because it is not well-defined. Humans as a group provide the standard for recognizing what is linguistically correct. Doing machine translation well is about convincing other humans that the job is being well done; human opinion as a whole sets the standard.

Can machines do language translation better than the average bilingual human?—We can ask this question. As soon as restrictions are placed on human performance, machine simulation might be sufficient to surpass the ability of any given individual human at any given time. The question of who is doing a better job continues to be decided by humans as a whole, but assuming the judging process involves a larger group of humans, or a more skilled human, then it still makes sense to say that a machine can outperform an average human.

How soon will truck drivers by replaced by machines?—We can ask this question. Machines do not need to emulate truck drivers to replace them. Machines might well be better at driving than the average truck driver. They might also be cheaper. Nevertheless, humanity continues to define what counts as ideal truck driving. At the extreme limits, it becomes difficult to formalize how exactly a vehicle should drive, and the issue reverts back to human opinion. For example, the issue of who should AI kill in a driverless car crash resists logical formalization because it interacts with human life. According to Goodall (2014), autonomous vehicles will certainly crash, some crashes will certainly involve a moral component, and there is "no obvious way to encode complex human morals effectively in software".

Do chess-playing programs play better chess than any human?—Yes, they do. For instance, the Komodo chess engine can reach an Elo rating of higher than 3300, which is about 450 points higher than any human currently playing chess (Regan 2014).

Are machines now as good as humans at playing chess?—This question has no utility because it is not well-defined. Chess-playing algorithms continue to be strengthened by humans, implying that humanity knows more about chess than any given machine. The point at which further strengthening becomes impossible is undecidable.

Does the human brain hold less than 500 exabytes of information?—This question has no utility because it is not well-defined. We don't have any means to formalize the representation of human abilities and thus we don't have any means to decide whether a not a given system which uses under 500 exabytes can emulate the behaviour of the human brain. Any question which presupposes a complete representation of the human mind in its entirety is a useless question.

Here we can see a pattern: the questions that seek to benchmark machine intelligence against human intelligence are useless, while the questions that consider how useful machines can be to humans are useful. Accordingly, we recommend that frameworks for evaluating the quality of machine "simulation" should be focused, not on the mimicry of human thinking, but on utility.

7. *Conclusion*

In recent years there have been suggestions of a possible future technological singularity, at which point computer programs would begin improving themselves recursively (e.g. Hutter 2012; Schmidhuber 2012). This concept, first identified by von Neumann (Stanislaw 1958), refers to the point at which machines start designing machines better than themselves, leading to a runaway effect, or intelligence explosion. According to this vision, smart machines will start designing successive generations of increasingly powerful minds, creating intelligence that far exceeds human intellectual capacity or control. Proponents perennially see the singularity as being 15 to 20 years off (e.g. Kurzweil 2005).

However, this concept is based on the flawed perspective of machines reaching a point where they emulate human intelligence. As we have seen, it is not possible to define or identify such a point. Given that the question of emulation is useless, then the idea of human labour being superseded is also meaningless. The rise of machine intelligence will not eliminate the value of human labour, but rather shift it away from repetitive formal tasks towards more complex psycho-social activities that are not as easily automated (see Turing 1951).

The focus in philosophy on directly contrasting human and machine intelligence has proved misguided. Over the past three decades a paradigm shift has occurred in the field of AI, with the focus moving away from a theory-driven quest to emulate the wholesale architecture of the mind, towards a data-driven approach which aims to achieve practical results in restricted domains (Cristianini 2014). A machine does not need to represent the full range of human abilities for it to be smart in some way. Human intelligence is social, embodied, and enactive, and very poorly described as symbol-processing or rule-following. AI, by contrast, aims to automate those aspects of human behaviour that are not unduly sophisticated.

AI is mostly developed and applied in limited domains, where computers' superior abilities in dealing with vast amounts of data quickly and following rules exactly are of greatest benefit. The performance of these systems is often already so far beyond human ability that comparing human and machine becomes wholly irrelevant. In other domains, poorer performance by machines will be tolerated as long as the "digital labour" is cheaper and more reliable. The question of who does the job better thus becomes moot.

Over the last 80 years, the process of computation defined by Church and Turing has proved extraordinarily useful to humans, and transformed modern society. In contrast, endless debates on whether the human mind can be matched or surpassed by AI are guaranteed to lead nowhere. Thus, the relevant questions for philosophy are not "is the mind a machine?" or "will there be a technological singularity", but rather, "how useful will machines be?" and "how will they change our lives?"

References

Aaronson, S. 2005. PHYS771 lecture 10.5: Penrose.

Aaronson, S. 2006. "Shtetl-optimized." The blog of Scott Aaronson. Reasons to believe.

Church, A. 1936. "An unsolvable problem of elementary number theory." *American journal of mathematics* 345–363.

Copeland, B. J. 2002. "The Church-Turing thesis." The Stanford Encyclopedia of Philosophy (Spring 2002 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2002/entries/church-turing/>

Copeland, B. J. 2003. "The Turing Test." In *The Turing Test*. New York: Springer: 1–21.

Cristianini, N. 2014. "On the current paradigm in artificial intelligence." *AI Communications* 27 (1): 37–43.

Cristianini, N. 2016. "A different way of thinking." *New Scientist* 232 (3101): 39–43.

Cuffaro, M. E., and Fletcher, S. C. (eds.). 2018. *Physical perspectives on computation, computational perspectives on physics*. Cambridge: Cambridge University Press.

Earman, J., and Norton, J. D. 1993. "Forever is a day: Supertasks in Pitowsky and Malament-Hogarth spacetimes." *Philosophy of Science* 60 (1): 22–42.

Fortnow, L. 2009. "The status of the p versus np problem." *Communications of the ACM,* 52 (9): 78–86.

French, R. M. 2012. "Moving beyond the Turing test." *Communications of the ACM,* 55 (12): 74–77.

Goodall, N. J. 2014. "Ethical decision making during automated vehicle crashes." *Transportation Research Record* 2424 (1): 58–65.

Gold, E. M. 1967. "Language identification in the limit." *Information and control* 10 (5): 447–474.

Harnad, S. 1992. "The Turing test is not a trick: Turing indistinguishability is a scientific criterion." *ACM SIGART Bulletin* 3 (4): 9–10.

Hayes, P. and Ford, K. 1995. "Turing Test considered harmful." In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. San Francisco: Morgam Kaufman Publishers: 972–977.

Hodges, A. 2009. *Alan Turing and the Turing test*. New York: Springer.

Hofstadter, D. H. 1980. *Godel, Escher, Bach: An eternal golden braid*. New York: Penguin Books.

Hutter, M. 2012. "Can intelligence explode?" *Journal of Consciousness Studies* 19 (1–2): 143–166.

Kieu, T. D. 2004. "Hypercomputation with quantum adiabatic processes." *Theoretical Computer Science* 317 (1–3): 93–104.

Kurzweil, R. 2005. *The singularity is near: When humans transcend biology*. New York: Penguin.

Legg, S. and Hutter, M. 2007. "Universal intelligence: A definition of machine intelligence." *Minds and Machines* 17 (4): 391–444.

Levesque, H. J. 2014. "On our best behaviour." *Artificial Intelligence* 212: 27–35.

Levin, L. A. 2003. "The tale of one-way functions." *Problems of Information Transmission* 39 (1): 92–103.

Li, M. and Vitányi, P. 2008. *An introduction to Kolmogorov complexity and its applications*. New York: Springer.

Longo, G. and Paul, T. 2011. "The mathematics of computing between logic and physics." B. Cooper and A. Sorbi (eds.). *In Computability in Context: Computation and Logic in the Real World*. London: Imperial College Press: 243–273.

Maguire, P. and Maguire, R. 2018. "On the measurability of measurement standards." *Croatian Journal of Philosophy* 19 (3): 403–416.

Maguire, P., Moser, P., and Maguire, R. 2015. "A clarification on Turing's test and its implications for machine intelligence." *Proceedings of the 11th International Conference on Cognitive Science*: 318–323.

McDermott, D. 2015. "What was Alan Turing's imitation game? Assessing the theory behind the movie." *The Critique*, January. URL: http://www.thecritique.com/articles/what-was-alan-turings-imitation-game/.

Müller, V. C. and Bostrom, N. 2016. "Future progress in artificial intelligence: A survey of expert opinion." In V. C. Müller (ed.). *Fundamental issues of artificial intelligence*. New York: Springer: 553–570.

Odifreddi, P. 1992. *Classical recursion theory: The theory of functions and sets of natural numbers*. Amsterdam: Elsevier.

Penrose, R. 1994. *Shadows of the mind*. Oxford: Oxford University Press.

Post, E. L. 1936. "Finite combinatory processes-formulation 1." *The Journal of Symbolic Logic* 1 (3): 103–105.

Regan, K. 2014. "The new chess world champion. Godel's Lost Letter and P=NP." Dec 28. URL: https://rjlipton.wordpress.com/2014/12/28/the-new-chess-world-champion/.

Rice, H. G. 1953. "Classes of recursively enumerable sets and their decision problems." *Transactions of the American Mathematical Society* 74 (2): 358–366.

Sandberg, A. and Bostrom, N. 2011. "Machine intelligence survey." *FHI Technical Report* 1.

Schmidhuber, J. 2012. "Philosophers and futurists, catch up! Response to the Singularity." *Journal of Consciousness Studies* 19 (1–2): 173–182.

Searle, J. R. 1980. "Minds, brains, and programs." *Behavioral and brain sciences* 3 (3): 417–424.

Shannon, C. E. and McCarthy, J. 1956. *Automata studies*. Princeton: Princeton University Press.

Shieber, S. M. 1994. "Lessons from a restricted Turing Test." URL: arXiv preprint cmplg/9404002.

Shieber, S. M. 2007. "The Turing Test as interactive proof." *Nous* 41 (4): 686–713.

Silver, D. et al. 2017. "Mastering the game of Go without human knowledge." *Nature* 550 (7676): 354.

Sloman, A. 2002. "The irrelevance of Turing machines to artificial intelligence." In M. Scheutz (ed.). *Computationalism: New Directions*. Cambridge: MIT Press: 87–127.

Turing, A. M. 1936. "On computable numbers, with an application to the Entscheidungsproblem." *Journal of Mathematics* 58 (345–363): 5.

Turing, A. M. 1947. *Lecture on the Automatic Computing Engine*. In Turing and Copeland 2004.

Turing, A. M. 1948. *Intelligent machinery*. In Turing and Copeland 2004.

Turing, A. M. 1950a. "Computing machinery and intelligence." *Mind* 59 (236): 433–460.

Turing, A. M. 1950b. Programmers. Handbook for Manchester Electronic Computer, University of Manchester Computing Laboratory. A digital facsimile of the original may be viewed in The Turing Archive for the History of Computing document. http://www.AlanTuring.net/programmers handbook.

Turing, A. M. 1951. *Intelligent machinery, a heretical theory*. In Turing and Copeland 2004.

Turing, A. M. 1954. *Solvable and unsolvable problems*. In Turing and Copeland 2004.

Turing, A. M. and Copeland, B. J. 2004. *The essential Turing: seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life, plus the secrets of Enigma*. Oxford: Clarendon Press Oxford.

Ulam, S. 1958. "John von Neumann 1903–1957." *Bulletin of the American mathematical society* 64 (3): 1–49.

Warwick, K. and Shah, H. 2015. "Can machines think? A report on Turing Test experiments at the Royal Society." *Journal of Experimental and Theoretical Artificial Intelligence* 28: 1–19.

Croatian Journal of Philosophy
Vol. XX, No. 58, 2020

Book Reviews

Leif Wenar, Blood Oil: Tyrants, Violence, and the Rules that Run the World*, New York: Oxford University Press, 2015, 552 pp.*

Oil is everywhere (xxxvi), in our clothes, cosmetics, roads, toys, electronics, household items. We use 1000 barrels of oil every second (xxxii). It is the most valuable traded commodity, worth more than 1 trillion dollars every year (xxxvi). Extracting, refining and selling oil is extremely profitable, and this business leads to large concentrations of power. The politics of dealing with oil and other highly valuable resources, such as gems, gas, minerals and diamonds, is thus sensitive and complex.

The analysis that Leif Wenar provides, building on rich literature and his previous work on *resource curse*, shows that trading with resources is intertwined with large amounts of unaccountable power which has imminent potential to destabilize the world—of which the Syrian refugee crisis is the most recent example (xliv). Countries that are rich with and dependent on natural resources and oil, such as Algeria, Angola, Sudan, Equatorial Guinea and many others, are at the same time countries with unstable governments and economies, suffering from conflicts, power abuse, corruption, authoritarianism and poverty. Research into international trade with resources, points to a role Western governments and their citizens have in bringing these "curses" on resource rich countries. *Blood Oil* is thus targeting primarily consumers from rich Western democracies that are largely unaware of origin of resources used to make everyday commodities.

Any time we fill up our cars, fly an airplane, or buy food and other products in our local stores, we might be sending some of our money to some of the worst dictators and strongmen in the world. This, as a result, brings enormous suffering to some of the poorest people, but also brings "curses" on us, manifested in terrorism, extremism, wars, climate change, economic crises and many other adversities (xxi). Political elites of countries rich in oil extract it and sell it on markets, avoiding any accountability to citizens of these countries; using that resource to strengthen their power; and not seldom, additionally oppress their subjects—as the cases of Theodoro Obiang of Equatorial Guinea and Saudi Arabia show. Once the global network of supply chains (xi), that connect our cell phones, laptops, cosmetics, jewelry or clothes to authoritarian regimes, conflict areas and countries hit by severe poverty is made obvious, a moral question arises: What should we do as consumers, knowing that our consumer choices affect poor living conditions of citizens of oil- and resource-exporting countries? This is

a question that Wenar addresses systematically and cautiously, not only providing an analysis of trade processes often-times hidden from us, but also proposing concrete courses of action that not only we as consumers, but also our governments and corporations should undertake. His book sets the stage for difficult policy changes that he claims must occur in the realm of oil and resource trade.

The book consists of 4 parts. The first and second part of the book trace causes of the resource curse and our contributions to it (xxvii). Furthermore, they show the effects of these curses and the complicated relationships and divisions that result from the unaccountable power derived from trade in oil and resources that affect everyone. Part three lays out basic principles on which change to the global trade market can be set forth. These principles pose a challenge to the current system of "coercion-based legal rights" (li), allowed by "might makes right" (xlv). The last part of the book is forward-looking, setting out policy proposals for a more just international trade praxis that should positively affect the countries struck by the resource curse and bring longer-term benefits to the rich importing countries and consumers.

The basic principle that Wenar uses in his analysis and policy proposals is "popular resource sovereignty". Popular resource sovereignty is a part of popular sovereignty (193), or the power of the people to freely determine their political status and pursue their development (196). Popular resource sovereignty, or right of peoples to their national resources and wealth, is codified in major human right conventions and thus recognized and ratified by most countries in the world. People have property rights over the natural resources of their country (203) and should be able to create and exercise laws that uphold that right.

However, as Wenar shows throughout the book, popular resource sovereignty is not at all the reality. For people to be able to authorize (or give tacit approval to) government or regime management of their resources, some minimal conditions must obtain (227–228): 1. Citizens need to have access to reliable, general information about the management of their resources; 2. They must not be subject to coercion, violence, brainwashing or extreme manipulation; 3. They must be able to deliberate and share information about resource management without fear of harm; and 4. They must be able to dissent to management of their resources without incurring severe costs. This translates to them having "at least bare-bones civil liberties and basic political rights" (228), that, without a doubt, many of the largest oil exporting countries do not provide for their subjects. Buying oil from these countries, where authorization of citizens is absent or highly unlikely, amounts to "carrying away stolen goods" (230). This theft is allowed, as Wenar stresses, by a fault in our international trade system—by a customary "might makes right", or the "effectiveness" rule.

It is this rule, a remnant from the old Westphalian era, that allows us to legally buy goods whose components are made from resources stolen from citizens by unaccountable regimes. Might translates into a legal right to sell resources (xlv). It is by this rule that "blood diamonds" were able to be legally sold on markets, or by which buying oil or resources from militias controlling some parts of territory, or authoritarians holding power over land and people, is recognized as lawful. This is a rule, Wenar shows, by

which our money is sent to coercive regimes and by which goods taken by force are declared legally clean (p. 122). This is a rule that allows violation of property rights, that legitimises and incentivises unaccountable power over territory and people, and that contributes to resource curse. Revealing of that rule is a basis for designing a change in trade system that Wenar dedicated a considerable amount of research to—as well as Thomas Pogge, who influenced his work.

From a moral standpoint, according to Wenar, there is no doubt that "might makes right" must be abandoned and international trade system reformed so as to recognize and respect popular resource sovereignty. However, difficulties are more than apparent, since oil business and trade in natural resources are highly lucrative and profit driven. Policy proposals should thus be carefully designed in order to address these obstacles. Principles of action for reform need to be strong enough to be recognized by the major global market players. Wenar believes that they are. Property rights are the pillars of free trade (266) while popular sovereignty, peace, human rights and rule of law (267) are principles already acknowledged by the majority of countries.

Clean Trade Policy proposal is perhaps the most important contribution of this book. The overall aim of these policies is to "end the global trade in stolen natural resources and to support public accountability over resources everywhere" (281). Wenar divides Clean Trade Policy into two parts. One set of policies is reserved for those countries where public accountability is severely lacking, while the other is aimed at countries where citizens have at least some degree of control over their national resources (283). For countries where minimal conditions of popular resource sovereignty do not obtain, importing countries can pass a Clean Trade Act, by which they can disengage commercially from unaccountable regimes, by making illegal the purchase of resources from these countries, and by denying entry into home jurisdiction and preventing any type of commercial and financial business with regime members and militants (284). This policy is "dramatic" (284), since regimes will retaliate in hope of protection of their interests. However, since the change applies solely to the laws of importing country and does not directly challenge legitimacy of foreign leaders and diplomatic recognition of resource-exporting countries, Wenar feels it is less dramatic than many other, familiar foreign policy options (285), such as sanctions. For this policy to be feasible, it is essential to establish reliable and bright-line standards for identifying countries where public accountability conditions are not met. Wenar proposes using already recognized indexes such as Freedom House report *Freedom in the World*, that ranks countries as *free*, *not free* and *partly free*, where *not free* oil exporting countries can be disqualified from trade. Clean Trade Acts need not immediately be passed for each exporting country where public accountability is lacking. To enable feasibility and avoid painful commercial shocks for importing countries, minimal steps may be taken, for example by first disqualifying "worst of the worst" (286). Clean Trade Act is thus used to enforce property rights of citizens of resource cursed countries and to stop dirtying hands of the consumers in importing countries.

Due to a realistic concern that many of the major players, such as China, will not block trade with these regimes, other measures are proposed to encourage trade partners to stop buying stolen resources (288–289). Apart

from various popular campaigns and boycotts (291–292) that consumers may engage in, countries that enforced Clean Trade Act may set up Clean Trade Trusts. These Trusts are bank accounts credited by the amount of money corresponding to the amount paid by other importing countries for natural resources *stolen* from the citizens of unaccountable exporting countries (290). This money is to be collected by tariffs on imported goods from countries that continue trading with regimes. It should be kept in the Trusts as a compensation to citizens whose property rights are being violated and returned once minimal conditions of accountability arise. This should ideally work as an incentive for trade partners to stop trading with these regimes and for citizens of unaccountable regimes to bring about positive changes in their home countries.

The other set of policies is targeting countries where citizens are at least partially free with the aim of supporting public accountability (321). Legal standards and sets of rules for companies operating both at home and abroad should be established. These would deal with the issues of bribery, corruption, money laundering, human rights violations, and lack of transparency (324). Various conditions could be built in the trading policies with the designated countries in order to reinforce their public accountability. One option is to introduce People's Funds or Sovereign Wealth Funds, which would accrue part of the profit from the oil revenue and distribute it directly to citizens in the form of "citizen share" (325–329).

Many concerns can be raised to some of these policies targeting efficiency, possible destabilizations of overall economy, possibility of violent regimes to retaliate or possibility of other unanticipated effects of these policies arising. Perhaps some additional concerns may be raised from standpoint of justice. If it is the citizens of these countries whom we should have to consider as recipients of remedy for violation of their property and human rights from which we all benefit, then Clean Trade Policy may be considered as too mild and too cautious of a proposal. Effects of the proposed policies seem very long-term and cannot be expected to ameliorate the circumstances of many individuals currently suffering under the regimes powered by trade in oil and resources. It is doubtful whether setting up Clean Trade Trusts is going to incentivise positive changes in resource-exporting countries, which are needed to bring about and elect more just governments. Therefore, many years may pass before a more accountable government is elected and the money collected from tariffs as compensation for property right violations is returned to the citizens. Additionally, in line with his laudable concern with feasibility, Wenar is primarily focused on proposing reforms based on internal policies of importing countries. These policies mirror his concern with non-intervention. By enforcing Clean Trade Policy importing country does not challenge political or diplomatic recognition of foreign regimes. It does not explicitly challenge the right of any regime to rule (285), no matter how it treats their subjects or whether it is democratically governed. It simply disengages from trading with these regimes. This caution can seem incompatible with human rights standards Wenar heavily leans on.

Clean Trade Policies may thus be supplemented by additional measures, such as sending material aid, investing in development projects or taking action where human right abuses are extremely severe. Another option is

promoting more open immigration policies in rich Western countries that benefit from the resource trade. This reform would acknowledge the more short-term considerations of concrete individuals that are owed duties of justice. Furthermore, it would sidestep direct intervention in the internal affairs of countries that severely violate human rights.

Many of the possible worries to these policies are raised and addressed in the book: worries about measures or standards proposed (293); interference in internal affairs of regimes (294–295); compatibility with WTO rules (297), some negative effects on countries banned from trade and on worst-off in both export and import countries (298–300); readiness of people for change (300–302); effects on energy supplies for importing countries; climate change (302–305), and others. *Beyond Blood Oil: Philosophy, Policy and The Future*, published in 2018 presents some additional criticism and answers provided by Wenar. Even with these issues taken into account, this book is a great contribution to the field of international resource trade. It systemises considerable body of literature and gives detailed analysis of the current praxis, with special consideration given to the contextualising of and to historical perspective on the issues. Wenar's writing is clear, revealing and accessible both to professionals and general public. His moral argument is compelling, inviting, and is built on widely shared values. More just international trade system is not merely an ideal, but the goal we should strive for and work on, as Wenar is doing—not just by his careful and precise writing, but also by other more practical activities he engages in.

TAMARA CRNKO
University of Rijeka, Rijeka, Croatia

Justin Garson, A Critical Overview of Biological Functions*, New York: Springer, 113 pp.*

In the book entitled *A Critical Overview of Biological Functions*, Justin Garson provides an accessible overview of the functions debate and delineates three canonical theories in the debate—the *selected effects theory*, the *fitness-contribution theory* and the *causal role theory*—and their specific ramifications, such as the *goal-contribution theory* and the *"weak" etiological theory*. In this critical overview, Garson also includes his preferred theory termed the *generalized selected effects theory*.

In the first chapter, entitled "What Is a Theory of Function Supposed to Do?", Garson emphasizes the important role that the notion of function plays in biology, philosophy, medicine, psychiatry, and ecology. An important philosophical task is to develop a theory which will best accommodate the notion of function in each of those disciplines. In line with this task, the author spells out three desiderata that every theory of biological function should satisfy. These desiderata are as follows: first, a theory should be able to distinguish a function of a trait from its accidental byproducts. For instance, "the function of my nose is to help me to breathe, but not to hold up my glasses, despite the fact that it does both and both are good for me, the latter is just a lucky accident." (4). Second, it should accommodate the explanatory dimension, i.e., "when we attribute a function to a trait, we purport to explain why the trait is there, that is, why organisms possess the

trait" (4). Third, the normative dimension of functional statements, that is, the logical possibility for a trait token to have a function that it cannot, in fact, perform (5). According to these desiderata, Garson evaluates prominent theories of biological functions.

In the second chapter, entitled "Goals and Functions", Garson starts with a historical overview of debates on the notions of purposefulness and goal-directedness related to the functioning of cybernetic machines in the 1920s and 1930s. He proceeds to the contemporary philosophical debates regarding biological functions that have started in the 1970s. In this chapter Garson provides an informative overview of theories preceding modern conceptualizations of biological functions, and lays out the foundation for following approaches, namely the selected effects theory, causal role theory, etc.

In the third chapter, entitled "Function and Selection", Garson examines selected effects theories. Here, the author shows how "the theory (*selected effects*) plausibly accounts for the explanatory and normative aspects of function" (33, italics added). The theory roughly states that a function of a trait is whatever it was selected for by natural selection or *some* natural process of selection. According to Garson, selected effects theory meets all three desiderata that the theory of functions should satisfy. Firstly, it can distinguish between a function and a lucky accident because a function of a trait is based on natural selection, hence it is not a mere accident. Secondly, this kind of theory provides an explanatory aspect of function because when one attributes a function to a trait, one offers a causal explanation for why the trait currently exists. Thirdly, a normative aspect of a function is met since a trait can *mal*function. In other words, it is possible for the trait not to perform its selected or "designed" function.

After laying out the main criticisms of the selected effects theory, Garson concludes this chapter with an exposition of his own preferred selected effects theory—the generalized selected effects theory. One of the important criticisms of the traditional selected effect views is that they do not apply to entities that do not reproduce. The generalized selected effects account can accommodate this problem. According to this view, entities can acquire functions in virtue of their differential persistence. To illustrate, Garson uses an example from neuroscience. He considers the formation of the mature synaptic structure of the human brain. Garson explains that formation of synapses and their pruning (which can be seen as a type of selection) can give rise to new functions in the brain even though there is no differential replication. According to Garson, the function of a trait consists in the activity that led to its differential *reinforcement* or its differential *reproduction* in a biological population. The first part of the definition of a generalized selected effects theory intends to cover various forms of processes of neural selection where there is no replication, and the second part of the definition covers the traditional part of the selected effects theory—natural selection (56–61). The third part of the definition, namely the one that refers to biological population, is meant to exclude some of the counterexamples for a selected effects theory (e.g. examples with clay crystals). Garson's own version of selected effects theory nicely addresses difficulties posed by critics towards the selected effects theory. By generalizing the definition, he tries to capture also the entities that do not reproduce, and by doing that, in a way, he advances the selected effect theory.

In the fourth chapter, entitled "Function and Fitness", Garson explains the fitness-contribution theory of function. He provides an overview of all the relevant theories that construe a function as a "contribution to the fitness of the organism that possesses it" (67). Some of the influential proponents of such a view are Christopher Boorse, Michael Ruse, and John Bigelow and Robert Pargetter. According to Garson, these theories can clearly meet only the first desideratum. We can distinguish between a function and an accidental effect since we can see the difference in the contribution of an effect on fitness (e.g. the function of the nose is to help us breathe and not hold up glasses because only the former effect is contributing to fitness, that is, it raises one's probability to survive and reproduce). However, Garson proceeds to claim that the second desideratum (the explanatory dimension) and the third desideratum (normativity) are not clearly met in the fitness-contribution theory of function.

In the fifth chapter, entitled "Function and Causal Roles", Garson discusses the causal role theories of biological functions. Garson explains: "According to this view, roughly, a function of a part of a system consists in its contribution to some system-level effect…" (81). The original causal role theory was developed by Robert Cummins. Cummins' causal role theory does not include a causal explanation of how a trait came about. For instance, it does not provide an explanation for the existence of a heart. Instead, causal role theory explains functions in terms of its contribution to a system in which it operates. Also, Cummins' view was further developed by Carl Craver and Paul Sheldon Davies. Their contribution to the development of the causal role theory includes utilizing the mechanistic framework to explain functions. Garson expounds two major problems for the causal role theory. The first problem is that the theory assigns a function to items that are intuitively non-functional. For instance, it is implausible to say that the function of a heart is to make beating sounds, but, proponents of the causal role theory must admit that in some contexts (depending on which effect of a trait we are interested in) this can be a function of the heart. The second problem is about distinguishing function and dysfunction. In some cases, causal role theory can ascribe a function to a trait that is clearly malfunctioning. For instance, if we are interested in how myelin degeneration causes paralysis, then on the present account, we would be forced to say that in this research context, myelin degeneration is functioning normally because it causes the effect under investigation (namely, paralysis).

Furthermore, Garson discusses function pluralism, which is motivated by the fact that biologists use both selected effects and causal role theories to assign functions to items, and, consequently, distinguishes two forms of pluralism. Function pluralism gained popularity due to its ability to capture different practices of ascribing functions. When biologists assign functions to items, in some cases they purport to causally explain why the item is there (selected effects theory), while in other cases, they purport to describe how the item contributes to a greater system (the causal role theory). Thus, according to pluralism, selected effects theory accommodates functions that are more prominent in evolutionary sciences (e.g. evolutionary biology) and the causal role theory captures functions in disciplines that do not rely on evolutionary explanations (e.g. physiology). This more "popular" version of pluralism Garson calls the *between-discipline* pluralism; different theories

of function are appropriate for different scientific disciplines. Garson also provides a new version of pluralism, the *within-discipline* pluralism. He emphasizes that it is possible that in one discipline scientists can use both theories in order to ascribe functions. For instance, even though a biologist does not explicitly appeal to selection when attributing functions to traits, she can do so implicitly. So, different concepts of a function can coexist within the same discipline, hence the name "within-discipline" pluralism.

In the sixth chapter, entitled "Alternative Accounts of Function", Garson expounds contemporary alternatives to classical theories of biological functions. Here Garson explains David Buller's "weak" etiological account, the family of systems-theoretic functions ("organizational view") and the modal theory of functions developed by Bence Nanay. Weak etiological theory defines function in terms of inheritance and past contribution of that function to fitness, thus, "a trait token in an organism has a function so long as that kind of trait contributed to the fitness of that organism's ancestor and it is inherited" (97). The family of systems-theoretic theories is "based on the idea that a trait token can acquire a function by virtue of the way that very token contributes to a complex, organized, system, and thereby to its own continued persistence, as a token." (97). The modal theory of functions says, roughly, that "the function of a trait token has to do with the behavior of that token in certain possible worlds." (97).

In the last chapter, entitled "Conclusion: What Next?", the author concludes the ideas developed in this book. Garson provides three main conclusions: (1) there are no viable alternatives to the selected effects theory since none other theory meets all desiderata; (2) if we accept pluralism it should be the "within-discipline" pluralism; and (3) he advocates his specific version of the selected effects theory—the generalized selected effects theory that is explained in the third chapter of the book.

To sum up, Garson's book provides a profound insight into the function debate. Through many informative examples, he illustrates and explains all relevant theories regarding biological function. In addition to explaining all three canonical theories and their misgivings, Garson also provides his own critical stance on the function debate, namely by introducing the generalized selected effects theory. His version of the selected effects theory is innovative in so far that it widens the scope of selected effects theory and, thus, provides new insights on the traditional debate. Garson's own approach belongs to the family of selected effects theories and, therefore, meets all the required desiderata that a biological function theory should meet. Furthermore, it should be emphasized that Garson introduces a new form of pluralism (the within-discipline pluralism) as a plausible position in the discussion about the nature of biological functions. Surely, this book provides a great impetus to philosophers and biologists to advance the debate on biological function.

VITO BALORDA*
University of Rijeka, Rijeka, Croatia

* This book review is an output of the "Theoretical Underpinnings of Molecular Biology" project (ThUMB) (IP-2018-01-3378) and doctoral grant (DOK-2018-09-7078) both financed by the Croatian Science Foundation.

Croatian Journal of Philosophy is published three times a year.
It publishes original scientific papers in the field of philosophy.

Croatian Journal of Philosophy is indexed in
The Philosopher's Index, PhilPapers, Scopus, ERIH PLUS
and in *Arts & Humanities Citation Index* (*Web of Science*).

Payment may be made by bank transfer

SWIFT PBZGHR2X
IBAN HR4723400091100096268

Croatian Journal of Philosophy is published with the support
of the Ministry of Science and Education
of the Republic of Croatia.

Instructions for Contributors

All submissions should be sent to the e-mail: cjp@ifzg.hr. Submissions must be in English and formatted to be double-spaced with suitably wide margins, an A4 page size, and automatic page numbering.

Articles are normally no longer than 8,000 words of main text including bibliography. The Journal will consider longer papers, but, once these limits are exceeded, authors should bear in mind the editorial policy that the acceptance bar raises with increasing length.

Manuscripts should be compiled in the following order: cover page; title; abstract (not exceeding 200 words); keywords (3 to 6); main text; appendices (as appropriate); references.

All the authors of a paper should include their full names, affiliations, postal addresses, telephone and fax numbers and email addresses on the cover page of the manuscript. If a paper is co-written, one author should be identified as the Corresponding Author. The cover page must be submitted as a separate document. All submitted manuscripts must be prepared for blind review, with revealing acknowledgements and self-identifying references removed.

Sources are cited in the text by the author's last name, the publication date of the work cited, and a page number if needed, e.g. (Barber 2007: 324). Full details appear in the reference list in which the year of publication appears immediately after the author's name:

Barber, A. 2007. "Linguistic Structure and the Brain." *Croatian Journal of Philosophy* 21 (7): 317–341.

Williamson, T. 2013. *Identity and Discrimination*. Oxford: Wiley-Blackwell.

The publication of a manuscript in the Croatian Journal of Philosophy is expected to follow standards of ethical behavior for all parties involved in the publishing process: authors, editors, and reviewers. The journal follows the principles of the Committee on Publication Ethics (https://publicationethics.org/resources/flowcharts).

ISSN 1333-1108
9 771333 110001