CROATIAN JOURNAL OF PHILOSOPHY

Vol. XXIII · No. 67 · 2023

Articles

Croatian Journal of Philosophy
Vol. XXIII, No. 67, 2023
https://doi.org/10.52685/cjp.23.67.1
Received: January 11, 2022
Accepted: December 13, 2022

How to Conquer the Liar and Enthrone the Logical Concept of Truth: an Informal Exposition

BORIS ČULINA
University of Applied Sciences Velika Gorica, Velika Gorica, Croatia

This article informally presents a solution to the paradoxes of truth and shows how the solution solves classical paradoxes (such as the original Liar) as well as the paradoxes that were invented as counterarguments for various proposed solutions ("the revenge of the Liar"). This solution complements the classical procedure of determining the truth values of sentences by its own failure and, when the procedure fails, through an appropriate semantic shift allows us to express the failure in a classical two-valued language. Formally speaking, the solution is a language with one meaning of symbols and two valuations of the truth values of sentences. The primary valuation is a classical valuation that is partial in the presence of the truth predicate. It enables us to determine the classical truth value of a sentence or leads to the failure of that determination. The language with the primary valuation is precisely the largest intrinsic fixed point of the strong Kleene three-valued semantics (LIFPSK3). The semantic shift that allows us to express the failure of the primary valuation is precisely the classical closure of LIFPSK3: it extends LIFPSK3 to a classical language in parts where LIFPSK3 is undetermined. Thus, this article provides an argumentation, which has not been present in contemporary debates so far, for the choice of LIFPSK3 and its classical closure as the right model for the truth predicate. In the end, an erroneous critique of Kripke-Feferman axiomatic theory of truth, which is present in contemporary literature, is pointed out.

Keywords: Paradoxes of truth; the truth predicate; the logical concept of truth; revenge of the Liar; the strong Kleene three-valued semantics; the largest intrinsic fixed point; Kripke-Feferman theory of truth.

1. *Introduction*

The concept of truth has various aspects and is a frequent subject of philosophical discussions. Philosophical theories usually consider the concept of truth from a wider perspective. They are concerned with questions such as—Is there any connection between the truth and the world? And, if there is—What is the nature of the connection?[1] Contrary to these theories, the analysis of the paradoxes of truth is of a logical nature because it deals with the internal semantic structure of a language, the mutual semantic connection of sentences, above all the connection of the sentences that speak about the truth of other sentences and the sentences whose truth they speak about. That is why every solution to the paradoxes of truth necessarily establishes a certain logical concept of truth.

The paradoxes of truth are "symptoms of disease" (Tarski 1969: 66): they show that there is a problem in our basic understanding of language, and they are a test for any proposed solution. Thereby, it is important to make a distinction between the *normative* and the *analytic* aspect of the solution.[2] The former tries to ensure that paradoxes will not emerge. The latter attempts to explain why paradoxes arise and to construct a solution based on that explanation. Of course, the practical aspect of the solution is also important. It tries to ensure a good framework for logical foundations of knowledge, for related problems in artificial intelligence and for the analysis of the natural language.

In the twentieth century, two solutions stood out, Tarski's (Tarski 1933, Tarski 1944) and Kripke's (Kripke 1975) solution. They initiated a whole series of considerations, from elaboration and critique of their solutions to proposals for different solutions. For the solution that is informally presented in this article, only Tarski's and Kripke's solutions are important, so other solutions will not be considered.[3]

Tarski's analysis emphasised the T-scheme as the basic intuitive principle for the truth predicate. According to Tarski, to examine the truth value of the sentence *"'snow is white' is a true sentence"*, we must examine whether snow is white. Thus, for the truth predicate the following must hold:

"snow is white" is a true sentence if and only if snow is white

This should be true for every declarative sentence S:

$\overline{S}$ is a true sentence if and only if S

where $\overline{S}$ is the name of the sentence S. For a particular sentence, we can always achieve this with quotation marks, as shown in the example of the sentence "snow is white". Tarski called this sentence scheme the

[1] A good overview of philosophical theories of truth can be found in (Glanzberg 2018). The author's position is set out in (Čulina 2020).

[2] In (Chihara 1979: 590), Chihara writes about "the preventative problem of the paradox" and about "the diagnostic problem of the paradox".

[3] An overview of various solutions can be found in (Beall et al. 2020).

T-scheme. However, if we apply the T-scheme to the sentence L: "$\overline{L}$ is a false sentence" (the famous Liar sentence), we will get a contradiction (the Liar paradox):

$\overline{L}$ is a true sentence if and only if $\overline{L}$ is a false sentence

Thus, Tarski's analysis showed the inconsistency of the T-scheme with the classical logic for the languages in which the Liar can be expressed, such as natural language.

Tarski's solution is to preserve the classical logic and to restrict the T-scheme to parts of the language. Tarski showed that if a language L meets some minimum requirements, we can consistently talk about the truth values of sentences of L only inside another "essentially richer" (Tarski's term) metalanguage ML. In ML, the T-scheme can only be set for the language L. This solution is in harmony with the idea of reflexivity of thinking and it has become very fertile for mathematics and science in general. For example, in chemistry, using the sentences of a language L we describe chemical processes, and using the sentences of ML we talk about the truth values of sentences of the language L.

Tarski does not deal with the analysis of the mechanism that leads to the paradoxes of truth, but only with the logical analysis of the formal inference of the contradiction. Not wanting to give up classical logic, the solution necessarily leads him to separate the metalanguage in which the T-scheme is expressed and the language for which the T-scheme is expressed, as a formal means of eliminating contradiction. Although Tarski does not explicitly say it anywhere, his solution suggests that the paradoxes of truth have their source in the violation of the reflexivity of thinking: talk about truth is an act of reflection whereby we transcend the original language. However, Tarski's solution is primarily of a normative nature. The mechanism of the paradoxes of truth is not analysed but paradoxes are blocked by a syntactic restriction. In ML we can speak only of the truth values of the sentences of the language L, so in ML the paradoxes of truth cannot be expressed at all. As for the liar paradox, the maximum approximation allowed by the syntactic restriction is the Limited Liar: when L is part of ML, under certain conditions, we can construct in ML the sentence

LL: $\overline{LL}$ is a false sentence of the language L[4]

If LL belonged to the language L, we could apply the T-scheme to LL:

$\overline{LL}$ is a true sentence of the language L if and only LL

According to the construction of the sentence LL, we get a contradiction:

$\overline{LL}$ is a true sentence of the language L if and only $\overline{LL}$ is a false sentence of the language L

[4] For example, this can be realized if for ML we choose the language of Peano's arithmetic and for L we choose Σ_n sentences of the language (Kaye 1991: 126).

However, from this contradiction it follows that *LL* does not belong to the language *L*. Thus, *LL* is certainly not a false sentence of the language *L*. So, it is a false sentence of the language *ML*.

Kripke showed that there is no natural syntactic restriction to the T-scheme as set out in Tarski's solution, but that we must look for the solution in the semantic structure of language. Consider the first example given by Kripke (Kripke 1975: 690). In the New Testament Saint Paul writes:

> One of Crete's own prophets has said it: "Cretans are always liars, evil brutes, idle bellies". He has surely told the truth.

In accordance with Tarski's approach, we can take as an object language the language composed of all the declarative sentences uttered by the Cretans together with the above statements of Saint Paul. In doing so, we will consider Saint Paul's first sentence to be true, which is an acceptable assumption. We will also assume that Saint Paul said all the above. For the sake of simpler expression, the sentence "Cretans are always liars, evil brutes, idle bellies" will be called "What one of Crete's own prophets said". The second St Paul's sentence is context dependent, so we will explicate it as the sentence "What one of Crete's own prophets said is true" and call it "What Saint Paul said". The application of the T-scheme for the object language gives us here:

1) What one of Crete's own prophets said is true if and only if Cretans are always liars, evil brutes, idle bellies
2) What Saint Paul said is true if and only if What one of Crete's own prophets said is true

According to 1), if What one of Crete's own prophets said is true then Cretans are always liars. So, What one of Crete's own prophets said is a lie. From this contradiction we conclude that What one of Crete's own prophets said is not true. By 2), we further conclude that What Saint Paul said is not true either. There is nothing paradoxical in the analysis so far.[5] However, let us consider what we can infer from the fact that What one of Crete's own prophets said is not true. By 1), it follows that Cretans are not always liars, evil brutes, idle bellies. So, we learned something about Cretans. It may seem odd that we have concluded something factual based on the T-scheme. However, we used Saint Paul's first statement as a factual assumption about the Cretans, and the T-scheme was only part of the logical mechanism by which we derived the above factual statement about the Cretans from this assumption. From a logical point of view, everything seems to be fine. However, we can imagine the extreme situation: that "one of Crete's own prophets" is the only Cretan, that he is not an evil brute or idle belly. That would mean he sometimes tells the truth. But we can go further and imagine that he made only one claim in his entire life—the

[5] Except perhaps for those who believe that everything written in the New Testament must be true.

one Saint Paul mentions. That would mean that What one of Crete's own prophets said is a true statement. And so, we got a contradiction again. In such a situation we are given a paradox: What one of Crete's own prophets said is true if and only if it is false, and so What Saint Paul said is true if and only if it is false.

In his article, Kripke describes a much more realistic situation in which the statements made have a certain truth value in normal conditions, but under some specific conditions they become paradoxical. In Kripke's words (Kripke 1975: 691):

> many, probably most, of our ordinary assertions about truth and falsity are liable, if the empirical facts are extremely unfavourable, to exhibit paradoxical features.

Kripke's analysis clearly showed that for a language in which one sentence speaks about the truth values of other sentences, what is expected and what is paradoxical in the language cannot be separated on the syntactic or internal semantic level: it depends on the reality that the language is talking about, and not on the way we use the language. Thus, according to Kripke, it is necessary to include this risk in the theory of truth. Sentences that speak of the truth values of other sentences, although syntactically correct and meaningful, under some conditions depending on the reality to which the language refers may not make a determinate claim about that reality: they will not give a classical truth value, *True* or *False*. Then we assign the third value to them: *Undetermined*. The meaning of the third value is simply that the sentence has no classical truth value. Such an analysis leads to the study of languages with partial two-valued semantics, which, by introducing *Undetermined* as the third value, is technically equivalent to the study of languages with three-valued semantics.

Kripke did not give any definite model. He gave a theoretical framework for investigations of various models—each fixed point in each monotone three-valued semantics can be a model for the truth predicate. Each such model gives a natural restriction on the T-scheme: the T-scheme is valid for all sentences that have a classical truth value in that model, while for the others it is undetermined.[6] However, as with Tarski, the proposed solutions are normative—we can express the paradoxical sentences, but we escape a contradiction by declaring them undetermined.

Kripke took some steps in the direction of finding an analytic solution. He preferred the strong Kleene three-valued semantics (SK3 semantics below) for which he wrote it was "appropriate" but did not explain why it was appropriate. One reason for such a choice is probably that Kripke finds paradoxical sentences meaningful. This elimi-

[6] For Kripke, as well as for my further analysis, the rules associated with the T-scheme are much more important than the T-scheme itself: that whenever the sentence S has a truth value, then the sentence "is true" has the same value and vice versa.

nates the weak Kleene three-valued semantics which in the standard interpretation[7] corresponds to the idea that paradoxical sentences are meaningless, and thus undetermined. Another reason could be that the SK3 semantics has the so-called investigative interpretation. According to this interpretation, this semantics corresponds to the classical determination of truth values, whereby all sentences that do not have an already determined value are temporarily considered undetermined. When we determine the truth values of some of these sentences, then we can also determine the truth values of some of the sentences that are composed of them, which were undetermined until then. For example, if we know that *S* is a true sentence and we do not yet know the truth of the sentence *T*, then according to the classical truth valuation of the conjunction, we do not yet know the truth of the sentence *S and T* (we will know it only when we know the truth value of the sentence *T*) but we do know that the disjunction *S or T* is true. This truth valuation corresponds exactly to the SK3 semantics.[8] Kripke supplemented this investigative interpretation with an intuition about learning the concept of truth in the presence of the truth predicate. That intuition deals with how we can teach someone who is a competent user of an initial language (without the truth predicate "to be true") to use sentences that contain the truth predicate. That person knows which sentences of the initial language are true and which are not. We give her the rule to assign the attribute "to be true" to the former and deny that attribute to the latter. In that way, some new sentences that contain the truth predicate, and which were undetermined until then, become determined. So, the person gets a new set of true and false sentences with which she continues the procedure. This intuition leads directly to the minimal fixed point of the SK3 semantics (MIFPSK3 below) as an analytically acceptable model for the concept of truth.

In the structure of the fixed points of a language with the truth predicate, two fixed points stand out, the minimal fixed point and the largest intrinsic fixed point. The first has the structural property that every sentence that has a classical truth value at the minimal fixed point has the same value at other fixed points. The largest intrinsic fixed point has the structural property that it is the largest fixed point such that every sentence that has a classical truth value in it has no opposite classical value at any other fixed point (it is compatible with all other fixed points). Kripke's work gives an internal characterisation of MIFPSK3, which follows from Kripke's description of the learning process of the concept of truth: at that fixed point, only those sentences whose truthfulness is based on the described learning process have a truth value. Starting with Kripke, the largest intrinsic fixed point is

[7] Some philosophers have given a different interpretation of the weak Kleene three-valued semantics, e.g. (Beall 2016).

[8] In weak Kleene three-valued semantics, if T would be a meaningless sentence, there is no need for further truth valuation, because automatically all sentences containing T are also meaningless.

mostly mentioned as an interesting solution because of its structural properties. Kripke writes (Kripke 1975: 709):

> The largest intrinsic fixed point is the unique "largest" interpretation of T(x) which is consistent with our intuitive idea of truth and makes no arbitrary choices in truth value assignments. It is thus an object of special theoretical interest as a model.

Since then, nothing much has changed in philosophical debates. Thus, forty years later, Horsten in his review article (Horsten 2015) writes:

> Until now, the intrinsic fixed points have not been investigated as intensively as they should perhaps be.

In (Čulina 2001) and in PhD thesis (Čulina 2004) I gave an analytic solution to the problem of the paradoxes of truth. In (Čulina 2001) it has been shown that this solution is precisely the largest intrinsic fixed point of the SK3 semantics (LIFPSK3 below) together with its classical closure. In this way, LIFPSK3 got a specific interpretation. This article provides an argumentation, which has not been present in contemporary philosophical discussions, for the choice of LIFPSK3 and its classical closure as the right model for the logical concept of truth. The solution will be informally described, and it will be demonstrated how it solves the classical paradoxes of truth (such as the original Liar) as well as the paradoxes that have been invented as counterarguments for various solutions to the paradoxes of truth ("the revenge of the Liar"). I will try to make the argumentation as simple as possible, so that the consideration can be followed by someone who does not have any special knowledge of the techniques related to Tarski's and Kripke's analysis. Finally, one of the confirmations of the naturalness of the solution of the problem of the logical concept of truth should be that such a solution can be explained in simple language, understood and used by any interested language user who does not have a special mathematical and philosophical education. All these informal considerations can be formalised by the means developed in (Čulina 2001). Some parts of the text are taken from (Čulina 2001) and PhD thesis (Čulina 2004). (Čulina 2001) contains the formal results to which this argumentation refers, while the PHD thesis contains the basic elements of the argumentation itself. However, much of what is only stated there has been elaborated and supplemented here to present rounded and convincing argumentation for the logical concept of truth introduced in these works.

2. *An analysis of the paradoxes of truth*

An analysis of the paradoxes of truth will be done on sentences. Tarski and Kripke state the technical reasons for this choice. In (Tarski 1944: 342) Tarski writes:

> By "sentence" we understand here what is usually meant in grammar by "declarative sentence"; as regards the term "proposition", its meaning is no-

> toriously a subject of lengthy disputations by various philosophers and logicians, and it seems never to have been made quite clear and unambiguous. For several reasons it appears most convenient to apply the term "true" to sentences, and we shall follow this course.

Kripke writes (Kripke 1975: 691):

> I have chosen to take sentences as the primary truth bearers not because I think that the objection that truth is primarily a property of propositions (or "statement") is irrelevant to serious work on truth or to the semantic paradoxes. On the contrary, I think that ultimately a careful treatment of the problem may well need to separate the "expresses" aspect (relating sentences to propositions) from the "truth" aspect (putatively applying to propositions). ... The main reason I apply the truth predicate directly to linguistic objects is that for such objects a mathematical theory of self-reference has been developed.

A convincing argument for choosing sentences for truth bearers was given by Quine in (Quine 1986: 1). This choice has an undoubted technical advantage because the subject of study is specific language forms, and not abstract objects of unclear nature. It is also a reflection of my deep conviction that language is not just a means of writing down and communicating thoughts but an essential part of thinking.[9]

Roughly, by "classical language" will be meant every language which is modelled upon the everyday language of declarative sentences. Due to definiteness, a language of the first order logic, which has an explicit and precise description of form and meaning, will be considered. By "language" will be meant an interpreted language, a language form together with an interpretation.

The interpretation of a first order language determines the external semantic structure of the language, a connection between the language and the subject matter of the language. The connection is based on external assumptions on the language use: (i) the language has its own domain of interpretation—a collection of objects that the language speaks of, (ii) every constant denotes an object, and every variable in a given valuation denotes an object, (iii) every function symbol symbolises a function which applied to objects gives an object, (iv) every predicate symbol symbolises a predicate which applied to objects gives a truth value, *True* or *False*. For simplicity, I will assume that the language has names for all objects in its domain. In doing so, $\overline{a}$ will be the name for an object a.

The inner organisation of a first-order language is determined by the rules of the construction of more complex language forms from simpler ones, starting with names, variables and function symbols for building terms, and with atomic sentences for building sentences. In these constructions we use special symbols which identify the type of the construction. With each construction, and thus the symbol of the construction, a semantic rule is associated that determines the semantic value

[9] My view of the essential role of language in thinking and rational cognition is explained in (Čulina 2021a).

of the constructed whole using the semantic values of the parts of the construction.[10] These rules determine the internal semantic structure of the language. The symbol of a language construction will be termed *logical symbol* or *logical constant* if the associated semantic rule is an internal language rule: the rule does not refer to the reality the language speaks of, except possibly referring to external assumptions of the language use. For example, connectives and quantifiers are logical symbols of the language.[11] I will further argue that the truth predicate "to be true" is also a logical symbol of the language. The interconnectedness of the truth values of sentences of a language belongs to one aspect of the concept of truth which I will term the *logical aspect of the concept of truth.* The connection of the truth value of a sentence with reality forms, for example, the second aspect of the concept of truth.[12] Since the paradoxes of truth occur in the context of the logical aspect of the concept of truth, I believe that each of their solutions establishes a certain logical concept of truth.

The external assumptions of the language use have grown from everyday use of language where we are accustomed to their fulfilment, but there are situations when they are not fulfilled. The Liar paradox and other paradoxes of truth are witnesses of such situations for the external assumption (iv). Let's consider the sentence L (the Liar):

L: $\overline{L}$ is a false sentence. (or "This sentence is false.")

Using the usual understanding of language, to investigate the truth value of L we must investigate what it says. But it says precisely about its own truth value, and in a contradictory way. If we assume it is true, then it is true what it says—that it is false. But if we assume it is false, then it is false what it says, that it is false, so it is true. Therefore, it is a self-contradictory sentence. What is disturbing is the paradoxical situation that we cannot determine its truth value.

The same paradoxicality, but without contradiction, emerges in the investigation of the following sentence I (the Truth-teller):

I: $\overline{I}$ is a true sentence. (or "This sentence is true.")

Contrary to the Liar to which we cannot associate any truth value, to this sentence we can associate the truth as well as the falsehood with equal mistrust. There are no additional specifications which would make a choice between the two possibilities.

I will begin the analysis of the paradoxes of truth with a basic observation that the above paradoxical sentences are meaningful because we understand well what they say, even more, we used that in the unsuccessful determination of their truth values. However, they witness

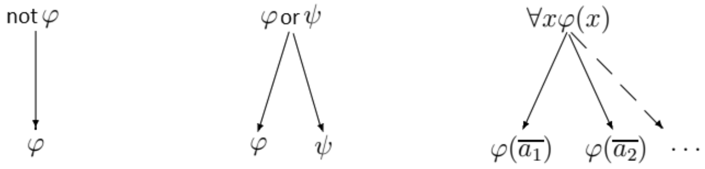
[10] In a given interpretation and a given valuation of variables, the semantic value of a term is the object described by the term and the semantic value of a sentence is its truth value.

[11] In (Čulina 2021b) the concept of logical symbol of a language is elaborated in more detail.

[12] In (Čulina 2020) various aspects of the concept of truth are analysed.

the failure of the classical procedure for the truth value determination in some "extreme" situations. According to the classical procedure, the examination of the truth value of a sentence is reduced to the examination of the truth values of the sentences from which it is constructed according to the classical truth value conditions for that type of construction. Thus, for example, the examination of the truth value of a sentence of the form φ *or* ψ is reduced to the examination of the truth values of the sentences φ and ψ. The reduction is performed according to the truth value conditions for the logical connective *or*: φ *or* ψ is true when at least one of the sentences φ and ψ is true, and false when both φ and ψ are false sentences. Likewise, a sentence of the form $\forall x\, P(x)$ (where $\forall$ is the standard symbol for *for all*) is a true when the sentences $P(\overline{a})$ are true for every object a from the domain of the language, and it is false when $P(\overline{a})$ is false for at least one object a. Thus, the examination of the truth value of a sentence comes down to the examination of the truth value of the sentences from which it is constructed (if these sentences contain free variables, then we must look at all valuations of these variables). Examining the truth values of these sentences is in the same way reduced to examining the truth values of the sentences from which they are constructed, etc.

We can visualise this procedure on the graph whose nodes are sentences of the language, where each sentence points with an arrow to the sentences to which, according to the classical truth value conditions, the examination of its truth value is reduced. Each type of sentence construction gives the corresponding type of elementary block of such a graph. To illustrate, the blocks corresponding to the constructions using negation (*not*), the disjunction (*or*), and the universal quantor ($\forall$) are shown below:



Each sentence has its own *semantic graph* to which the sentence is a distinguished node, and the graph is composed of all sentences on which, according to the truth value conditions, the truth value of a given sentence hereditarily depends.[13]

To determine the truth value of a given sentence, according to the classical truth value conditions, we must investigate the truth values of all sentences to which it points, then possibly, for the same reasons, the truth values of the sentences to which these sentences point, and so on. Every such path along the arrows of the graph leads to atomic sen-

[13] The semantic graph of the whole language can be defined analogously. The semantic graphs of individual sentences are its subgraphs.

tences (because the complexity of sentences decreases along the path). In situations where a language doesn't talk about the truth values of its own sentences, the truth values of its atomic sentences don't depend on the truth values of some other sentences. The atomic sentences are the leaves of the semantic graph of the given sentence. To investigate their truth values, we must investigate external reality they are talking about. The classical assumption of a language is that every atomic sentence has a definite truth value. So, the procedure of determination of the truth value of the given sentence stops and gives a definite truth value, *True* or *False*. Formally, this is secured by the recursion principle which says that there is a unique function from sentences to truth values, which obeys the classical truth value conditions and its values on atomic sentences are identical to externally given truth values.[14] Such is, for example, the language of a scientific field, but not the everyday language in which there are frequent discussions about the truthfulness of claims made by others. In such situations, the above analysis can be, and is, disrupted when atomic sentences use the truth predicate to speak of the truth values of other sentences of the language. These are sentences of the form $T(\overline{\varphi})$, where "T" is the symbol for the truth predicate "to be true", and $\overline{\varphi}$ is the name of a sentence φ of the language. Such an atomic sentence is not a leaf of a semantic graph, but points with an arrow to the sentence φ on which its truth value depends:

$$\begin{tikzcd}
T(\overline{\varphi}) \arrow[d] \\
\varphi
\end{tikzcd}$$

The truth value conditions of this construction are the basic conditions of the logical concept of truth: that $\mathrm{T}(\overline{\varphi})$ is true when φ is true, and $\mathrm{T}(\overline{\varphi})$ is false when φ is false. Given that the semantic rule of this construction is an internal semantic rule (it connects the truths of sentences $\mathrm{T}(\overline{\varphi})$ and φ independently of the reality the language speaks of), the truth predicate is a logical symbol of the language, in the same way that, for example, connectives and quantifiers are logical symbols of the language. In this sense, it is perfectly correct to speak of this concept of truth as a *logical concept of truth*. The only difference in relation to connectives and quantifiers is in universality. Only a language that has its own sentences in the domain of its interpretation (possibly through coding) can have a logical symbol of its own truth predicate.

[14] Note that, even when we know the true values of the leaves, this procedure is generally not computable because although the semantic graph of a given sentence has finite depth (the reduction to the leaves takes place in the finite number of steps), the leaves themselves can be infinitely many.

In the presence of the truth predicate, it can happen that the procedure of determination of the truth value of a given sentence does not stop at atomic sentences but, under the truth value conditions of the truth predicate, continues through each atomic sentence of the form T($\overline{\varphi}$) to the sentence φ. Because of the possible "circulations" or other kinds of infinite paths, there is nothing to ensure the success of the procedure. Truth paradoxes just witness such situations. Five illustrative examples follow.

$$\begin{tikzcd} \mathrm{T}\left(\overline{1+1=3}\right) \arrow[d] \\ 1+1=3 \end{tikzcd}$$

The procedure of the truth value determination has stopped on the atomic sentence for which we know is false, so T($\overline{1+1=3}$) is false, too.

The Liar: For L: T($\overline{\textit{not } L}$) we have

$$\begin{tikzcd} L\colon \mathrm{T}\left(\overline{\textit{not } L}\right) \arrow[d] \\ \textit{not } L \arrow[u, bend left=60] \end{tikzcd}$$

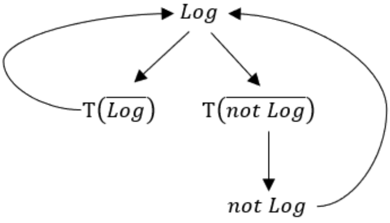
But now the procedure of the truth value determination has failed because the truth value conditions can't be fulfilled. The truth value of T($\overline{\textit{not } L}$) depends on the truth value of *not* L and this again on L: T($\overline{\textit{not } L}$) in a way which is impossible to obey.

The Truth-teller: For I: T($\overline{I}$) we have

$$\begin{tikzcd} I\colon \mathrm{T}\left(\overline{I}\right) \arrow[loop left] \end{tikzcd}$$

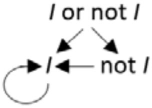
Now, there are, as we have already seen, two possible assignments of the truth values to the sentence I. But this multiple fulfilment we must consider as a failure of the classical procedure, too, because the procedure assumes to establish a unique truth value for every sentence.

The Logician: *Log*: $\mathrm{T}(\overline{Log})$ or $\mathrm{T}(\overline{not\ Log})$ (This sentence is true or false)


Log
$\mathrm{T}(\overline{Log})$
$\mathrm{T}(\overline{not\ Log})$
not Log

If *Log* were false then, by the truth conditions, $\mathrm{T}(\overline{not\ Log})$ would be false, not *Log* would be false too, and finally *Log* would be true. Therefore, such valuation of the graph is impossible. But if we assume that *Log* is true, the truth conditions generate a unique consistent valuation. Therefore, the truth determination procedure gives the unique answer—that *Log* is true.

The law of excluded middle for the Truth-teller: *I or not I*



Now there are two truth valuations of the semantic graph of the sentence *I or not I*. In both valuations, it takes the same value: *True*. However, in one valuation the sentence *I* takes the value *True* and in the other *False*. Given that the classical procedure requires that not only the initial sentence, but every sentence included in the examination, if it has a truth value, then has a unique truth value, we must also consider this situation as a failure of the classical procedure for determining the truth value of the sentence *I or not I*. Having failed to determine the truth value of sentence *I*, we have not been able to determine the truth value of the sentence *not I*, and therefore neither of the sentence *I or not I*.

The paradoxes of truth emerge from a confrontation of the implicit assumption of the success of the classical procedure of the truth value determination and the discovery of the failure. As previous examples show such assumption is an unjustified generalisation from common situations to all situations. We can preserve the classical procedure, also the internal semantic structure of the language, but we must reject universality of the assumption of its success. The awareness of that transforms paradoxes to normal situations inherent to the classical procedure. I consider this the diagnosis of the paradoxes of truth.

3. *The proposed solution*

The previous diagnosis shows us the way to the solution—the formulation of the *partial two-valued semantics* of language which, when the procedure of determining the truth value of a given sentence gives a unique truth value, *True* or *False*, attaches that value to the sentence, and when the procedure fails, it does not attach any truth value to the sentence. This kind of semantics can be described as the *three-valued semantics* of language—simply the failure of the procedure will be declared as the third value (*Undetermined*). It has not any additional philosophical charge. It is only a convenient technical tool for the description. In formulating the partial two-valued semantics, we will start from these properties:

1) The semantics coincides with the classical semantics on atomic sentences whose truth values are determined by the external reality they are talking about.
2) In the semantics all sentences are meaningful.
3) The semantics has classical truth value conditions for connectives and quantifiers.
4) In the semantics $\mathrm{T}(\overline{\varphi})$ is true when φ is true, and false when φ is false (a variant of the T-scheme).
5) When the classical procedure of determining the truth value of a given sentence assigns it a unique truth value then the semantics assigns that value to the sentence, otherwise it does not assign a truth value to the sentence.

Properties 1) and 4) need no comment. Property 2) was commented at the beginning of this analysis. The fact that we cannot determine the truth values of paradoxical sentences does not mean that they are not meaningful. We understand their meaning quite well. Moreover, we use this meaning essentially in the (unsuccessful) determination of their truth values. The consequence of this property is that all sentences have meaning, regardless of whether some part of the sentence is paradoxical or not. Otherwise, as soon as one part of the sentence was paradoxical, the whole sentence would be meaningless.[15] Here is one argument as to why it is not an acceptable solution to consider paradoxical sentences to be meaningless. If we were to accept that paradoxical sentences have no meaning, it would make no sense to determine their truth values. Thus, we could not determine which sentences are paradoxical, i.e., they have no meaning.[16]

For property 3) it is only important to note that the rejection of the success of the classical procedure of the truth value determination doesn't change the meaning of the classical truth value conditions. They are stated in a way independent of the assumption that sentences

[15] This would lead to the weak Kleene three-valued semantics of the language.

[16] Thus this argument rejects the weak Kleene three-valued semantics as a solution to the paradoxes of truth.

must have a truth value. They specify the truth value of a compound sentence in terms of the truth values of its direct components regardless of whether they have truth values or not. The lack of some truth value may lead, but does not have to, to the lack of the truth value of the compound sentence. For example, the truth value conditions of the sentence φ *and* ψ are: φ *and* ψ is true when both φ and ψ are true, and false when at least one of the sentences φ and ψ is false. It says nothing about the existence of the truth values of φ and ψ, but only sets conditions among the truth values. The functioning of the truth value conditions in the new situation is illustrated by the example of the following sentences (where L is the Liar, and I is the Truth-teller):

$$L \textit{ or } 0 = 0$$

By the classical truth value conditions for the connective *or*, this sentence is true precisely when at least one of the basic sentences is true. Because 0=0 is true consequently the total sentence is true, although L has not a truth value. Equally, if we apply the truth value conditions on the connective *and* to the sentence

$$L \textit{ and } 0 = 0$$

the truth value will not be determined. Namely, for the sentence to be true both basic sentences must be true, and it is not fulfilled. For it to be false at least one basic sentence must be false, and this also is not fulfilled. So, non-existence of the truth value for L leads to non-existence of the truth value for the whole sentence. Let's analyse

$$I \textit{ or not } I$$

Since I does not have a truth value, *not I* does not have a truth value, so *I or not I* also does not have a truth value.

Property 5) is a key property. It expresses the basic idea of this approach: the lesson of the paradoxes of truth is that the classical procedure of determining the truth value does not have to succeed. By failure we mean that, respecting the classical conditions of truth, we cannot assign a truth value to a sentence, or we can assign two truth values to it. However, property 5) stated in this way is not precise enough because the classical determination of truth values is not an algorithmic process and in concrete situations we manage to implement it in various ways. Furthermore, rejecting the assumption of the existence of a unique truth value complicates the process, because now it is not enough to find one valuation of a given sentence, but it is necessary to examine whether there are other valuations, not only of the given sentence but also of other sentences included in the examination. That's why we must give property 5) a more objective formulation that does not talk about the real or idealised process of determining the truth values of sentences, but about the existence of these values. In (Čulina 2001) it is shown that:

There is a unique partial two-valued semantics (association of truth values to sentences) with the following properties:

1) The semantics coincides with the classical semantics on atomic sentences whose truth values are determined by the external reality they are talking about.
2) The semantics has classical truth value conditions for connectives and quantifiers.
3) In the semantics $\mathtt{T}(\overline{\varphi})$ is true when φ is true, and false when is false (a variant of the T-scheme).
4) The semantics is unique in the sense that on the set of all sentences to which it associates truth values, every other semantics that fulfils the previous three conditions does not associate different truth values (it can happen that the other semantics does not associate a truth value to some of these sentences).
5) The semantics is the largest such semantics in the sense that on the set of sentences to which it does not assign truth values, every other semantics that fulfils the previous four conditions also does not assign truth values.

Below, I will call this semantics *the partial two-valued semantics*. This is exactly the requested extension of classical semantics to situations where it is not guaranteed that every sentence is true or false because this semantics accurately identifies when the classical procedure of determining the truth value of a sentence will succeed and when it will not.

This result gives a "license" to the classic procedure of determining the truth value of a sentence in situations where not all sentences have a truth value. The restriction of the partial two-valued semantics to the semantic graph of a given sentence is the truth valuation of the graph, which the classical procedure should determine by its success or failure. Thereby, the classical procedure does not need to determine the entire valuation, but only that part that is sufficient to determine the truth value of a given sentence or to determine that it has no truth value. For example, to determine the truth value of the sentence $\exists x\, \varphi(x)$ (where $\exists$ is the standard symbol for exists), if among all sentences of the form $\varphi(\overline{a})$ we find one that is true then we do not have to examine the others, nor do we have to worry about whether any of them is undetermined. Likewise, when we know that some sentences are undetermined, we can use this in determining non-existence of the truth values of other sentences. For example, for the sentence L *and* $0{=}0$, knowing that L is undetermined allowed us to conclude that L *and* $0{=}0$ is also undetermined. Thereby, not only the truth value conditions for the connective *and* do not give us the truth value for L *and* $0{=}0$ but the failure of the classical procedure in determining the truth value of L leads to the failure of determining the truth value of L *and* $0{=}0$. This example shows that not only the classical truth value conditions of the conjunction of two sentences do not depend on whether these sentences have a truth value but the conditions also determine how the failure of the determination of truth values is propagated. It is easy to see that this is also true in general: all the truth value conditions not only determine the connection between truth values but also determine how the failure of the determination is propagated. If we look at the associated three-valued semantics, it is not difficult to show that these are precisely the conditions of the SK3 semantics. Thus, SK3 have a special interpretation here: the SK3 conditions are the classical truth value conditions supplemented by the conditions of propagation of the failure to determine truth values.

In (Čulina 2001) it was proved that in the labyrinth of literature on the paradoxes of truth (Beall et al. 2020), the partial two-valued semantics described above is positioned as the largest intrinsic fixed point of the SK3 semantics (LIFPSK3) with a specific interpretation. In that way, the above presented argumentation for the partial two-valued semantics is also the argumentation for the choice of LIFPSK3 among all fixed points of all monotone three-valued semantics for the right model of the logical concept of truth.

In (Kremer 1988: 245), Kremer writes:

> Within Kripke's theoretical framework there are two leading candidates for the "correct" interpretation of the truth predicate: the minimal fixed point and the largest intrinsic fixed point. … We are thus led to distinguish two plausible versions of the principle of the supervenience of semantics. First, there is the view of that the correct interpretation of truth is the minimal fixed point; as we saw, this has often been taken to be "Kripke's theory of truth". Second, there is the view that the largest intrinsic fixed point is the correct interpretation of truth. Unfortunately for the champion of supervenience, there seem to be considerations in support of both of these views.

I will give some arguments as to why I consider LIFPSK3 with the interpretation described in this article to be a better solution than MIFPSK3 with Kripke's interpretation. The main argument concerns the content-wise interpretations of these fixed points. In Kripke, it is an interpretation of learning the concept of truth in the presence of the truth predicate, here an interpretation of determining truth values of sentences that language users actually do.

In Kripke, the SK3 semantics has an investigative interpretation: while we have not yet determined the truth values of some sentences, they are undetermined. In the process of learning the concept of truth in the presence of the truth predicate, more and more sentences gain truth value. So, some hitherto undetermined sentences become determined, which, according to the truth value conditions, entails that some others sentences become determined. However, some sentences will remain undetermined forever. Thus, as Visser noted in (Visser 1989: 651), the SK3 interpretation changes: "not yet" interpretation of undetermined value in the learning process (because we haven't learned the concept of truth enough yet), in MIFPSK3 becomes "not ever" interpretation (a sentence is undetermined because its truth value can never be determined through the process of learning the concept of truth). In the interpretation developed in this article, undetermined sentences are those sentences to which the classical procedure of determining truth values does not give a unique truth value. SK3 naturally derives from the classical procedure of determining truth values, which in the presence of the truth predicate is not always successful. In this interpretation, SK3 is simply the classical semantics complemented by the propagation of its own failure. For example, let's analyse the Liar in both interpretations. In Kripke's interpretation, learning the predicate of truth, we will not face the Liar at any level to determine its truth value. The Liar is simply inaccessible to us in that process and, if we strictly adhere to the metaphor of learning the predicate of truth, we will never even know that the Liar is inaccessible to us. In contrast, the interpretation developed in this article provides only a formal framework for the process a language user actually undertakes when encountering the Liar and examining its truth value. The language user will easily determine that the Liar is undetermined.

Furthermore, in Kripke's interpretation, language users learn the truth predicate in an extensional way, collecting more and more sentences that fall under the predicate and more and more sentences that do not fall under the predicate. However, as explained in this article, the truth predicate is a logical concept: it is determined by the internal semantics of language and it should not be learned experientially, just as, for example, the logical meaning of the connective *and* should not be learned experientially. As we know the meaning of the connective *and* when we are given its truth value conditions, so we know the meaning of the truth predicate, when we are given its truth value conditions: $\mathrm{T}(\overline{\varphi})$ is true when φ is true, and it is false when φ is false. From this definition of the logical concept of the truth predicate, which is a variant of the T-scheme, and which corresponds to the basic intuition of the language user, arises the interpretation developed in this article which gives LIFPSK3. In Kripke's case the opposite is true: from the intuition about learning the truth predicate follows MIFPSK3 as an extensional a posteriori definition of the truth predicate.

Finally, to learn someone which sentences to associate with the predicate "is true", we must first know it ourselves. According to Kripke's interpretation, someone should first learn us which sentences to associate with the predicate "is true". Thus, the idea of learning the truth predicate leads to an infinite regress: learning grounded truth is ungrounded.

That the aspect of learning the concept of truth and understanding the concept of truth is not one and the same, Yablo has already noted in (Yablo 1982: 118), but in the context of MIFPSK3:

> If the inheritance aspect is the one lying behind the attempt to picture grounding in terms of the learning of 'true', then the dependence aspect is the one behind the attempt to picture grounding in terms of the understanding of 'true'. What do we do when we have to evaluate a sentence—say "The

> sentence 'Snow isn't white' is true" or the sentence "The sentence 'Snow is white' is true' is not true"—involving complicated attributions of truth? Evidently, we try to figure out what its truth-value depends on, and then what that depends on, and so on and so forth in the hope of eventually making our way down to sentences not containing 'true' which can be evaluated by conventional means. … But the fact that the majority of those who grappled with grounding before Kripke tended to see things from the standpoint of dependence suggests that there is something intuitively satisfying about the dependence approach.

As already commented, MIFPSK3 and LIFPSK3 have distinguished structural properties in the structure of all fixed points. Kripke's description of the learning process gave a characterisation of MIFPSK3 independent of other fixed points. The analysis developed in this article provides a characterisation of LIFPSK3 that is also independent of other fixed points. However, while Kripke's characterisation is global—the learning process yields all the truths and falsehoods of MIFPSK3—the LIFPSK3 characterisation developed here is local: the truth value determination of a given sentence takes place only on the semantic graph of the sentence. The characterisation of LIFPSK3 developed here, and not the Kripke's characterisation of MIFPSK3, corresponds to the way a language user determines the truth value of a sentence. Starting from a given sentence, the language user tries to determine its truth value by examining its semantic graph, and not by collecting more and more true and false sentences according to the instructions for learning, and hoping that the given sentence will appear in one of those groups. In this letter case, as it was illustrated above on the example of the Liar, the language user can never determine that a sentence is undetermined: it is constantly in the "not yet" interpretation and can never switch to the "not ever" interpretation.

LIFPSK3 contains MIFPSK3 as a subset, which can also be considered an advantage of LIFPSK3. To all sentences, which MIFPSK3 assigns a truth value, LIFPSK3 also assigns this same value. But in addition, LIFPSK3 assigns truth values to sentences that are undefined in MIFPSK3. Such, for example, is the sentence the Logician, which is assigned the value *True* by the classical truth determination procedure, as shown above, while it is undetermined in MIFPSK3. Of course, a remark can be made here that MIFPSK3 is a better choice for this very reason, because in MIFPSK3 truth values are given only to those sentences whose truth value is "grounded" in the reality that the language speaks about. Such is not, for example, the Logician, but the sentence *L or* $0 = 0$ is: this sentence is true because the atomic sentence $0 = 0$ is true. Although Kripke formally calls grounded all sentences that have a truth value in MIFPSK3, on an intuitive level they are grounded because their truth values are determined by the process of learning the concept of truth which starts from the truth values of atomic sentences that speak of reality. However, as already stated, paradoxes of truth fall under the logical concept of truth that connect the truth values of sentences independently of the reality that the language speaks of, so the solution to the paradoxes should not include reality. The classical procedure for determining truth values does not require us to examine whether a sentence is grounded or not, but only whether we can associate a unique truth value with it or not. Of course, in this examination, we can arrive at atomic sentences that talk about external reality, but we don't have to, as the example of the Logician shows. In this comparison it is also seen that the choice between Kripke's interpretation of MIFPSK3 and the interpretation of LIFPSK3 described in this paper is a choice between non-logical and logical concept of truth.

The next section will show that some of Gupta's critiques (Gupta 1982) of fixed points apply to MIFPSK3 but not to LIFPSK3. Thus, the critiques turn into the argument that LIFPSK3 is a more acceptable model for the truth predicate than MIFPSK3.

So, for now we have two semantics of a language with the truth predicate. We have the *classical* or *naive semantics* in which paradoxes occur because this semantics assumes that each sentence is true or false, i.e., it assumes that the process of determining truth values always gives an unambiguous answer. And we have its repair to the two-valued partial semantics of the language, i.e., to the three-valued semantics of the language, which accepts the possibility of failure of the classical procedure of determining truth values. I will call this semantics the *primary semantics* of the language. However, to remain on the partial two-valued semantics would mean that the logic would not be classical, the one we are accustomed to. Concerning the truth predicate itself, it would imply the preservation of its classical logical sense in the two-valued part of the language extended by the "silence" in the part where the classical procedure fails. For example, the T-scheme is true only for sentences that have a classical truth value. For other sentences it is undetermined. Although in a meta-description, $\mathrm{T}(\overline{\varphi})$ has the same truth value (in the three-valued semantic frame) as φ, that semantics is no longer the initial classical semantics (although it extends it) nor it can be expressed in the language itself: the language is silent about the third value. Or better said, the third value is the reflection in a metalanguage of the silence in the language. So, the expressive power of the language is weak. For example, the Liar is undetermined. Although we have easily said it in the metalanguage, we cannot express in the language itself, because, as it has already been said (in the metalanguage), the Liar is undetermined. Not only that this "zone of silence" is unsatisfactory for the above reasons (it leads to the three-valued logic, it loses the primary sense of the truth predicate and it weakens the expressive power of the language), but it can be overcome by a natural *additional valuation* of the sentences which emerges from recognising the failure of the classical procedure. "Natural", in the sense that it is precisely this move that a language user makes in the end when faced with the failure of the classical procedure. This point will be illustrated on the example of the Liar. On the intuitive level of thinking, by recognising the Liar is not true nor false we state that it is undetermined. However, this is not a claim of the original language but of the metalanguage in which we describe what happened in the language. Moreover, in the metalanguage, we can continue to think. Since the Liar is undetermined, it is not true what it claims—that it is false. Therefore, the Liar is false. But this does not lead to restoring of the contradiction because by moving to the metalanguage we have made a *semantic shift* from the primary partial two-valued semantics (or the three-valued semantics) toward its two-valued meta description. Namely, the Liar talks of its own truth in the frame of the primary semantics, while the last valuation is in the frame of another semantics, which I will term the *final semantics* of the language. The falsehood of the Liar in the final semantics doesn't mean that it is true what it says (that it is false) because the semantic frame is not the same. The falsehood of the Liar in the final semantics means that it is false (in the final semantics) what the Liar talks of its own primary semantics: that it is false in the primary semantics (because it is undetermined in the primary semantics). So, not only have we gained a contradiction in the naive semantics, i.e., the third value in the primary semantics, but we also have gained additional information about the Liar.

A key element in the above consideration is a semantic shift in thinking. It is closely related to the reflexivity of thinking, which appears in two variants in the paradoxes of truth. The first variant takes place in the primary semantics of the language, and the second in the transition to the final semantics of the language. In the first variant, the reflexivity of thinking occurs in the transition from the use of the sentence S to the mention of the sentence S. The most significant example of this transition in the context of the paradoxes of truth is the transition from the statement S to the statement "$\overline{S}$ is a true sentence". Thereby, two aspects of the concept of truth should be distinguished. To examine whether "snow is white" is a true sentence, we must investigate reality, see what colour the snow is. So, the truth value of that sentence depends on reality. Therefore this aspect of the concept of truth is not of a logical nature. To examine the truth value of the sentence ""snow is white" is a true sentence" we must examine the truth value of the sentence "snow is white". Thus, the truth value of that sentence also depends (indirectly) on reality. However, the truth predicate only articulates this transfer of truth, just as, for example, the connective *and* articulates the transfer of the truth values of a conjunction to the conjuncts. According to the classical truth conditions on the truth predicate, the predicate connects the truth values of two sentences in the primary semantics in a way that is independent of the reality the language is talking about. That is why the truth predicate is a logical symbol in the primary semantics and falls under the logical notion of truth in the primary semantics.

Another variant of reflexivity of thinking occurs at the level of the whole language—in the transition from the original language to the metalanguage by which we describe the original language. In the context of the paradoxes of truth, this transition was illustrated above on the example of the Liar, when we concluded that the Liar is undetermined in the primary semantics. This conclusion belongs to the metalanguage by which we describe truth valuations in the original language. The metalanguage has the same syntax as the original language (we will see below that the predicate "is undetermined" can be defined by the predicate "is true"), but not the same semantics: it has a different connection between the truth values of sentences, and this connection is a classical two-valued semantics, called here the final semantics of the language. The metalanguage is a classical language with classical semantics and an external reality that it talks about in the same way that a classical language is, for example, the language we use to describe car engines. Only here it is not so obvious, because the external reality that the metalanguage talks about is another language that has the same syntax as the metalanguage (but not the same semantics). The key element of the semantic distinction between these two languages is the truth predicate that we have analysed so far. Given that we now have two languages, we must first express this predicate of truth more precisely, to make it clear that it is the truth predicate of the original language. That is why instead of "is true" we will now use "is true in the primary semantics". This does not change its role in primary semantics one bit—the logical role described above. In the final semantics, the sentence "$\overline{S}$ is true in the primary semantics" has the same meaning as in the primary semantics: it asserts that the sentence S is true in the primary semantics. However, in the final semantics the truth predicate of the primary semantics connects the truth value of the sentence S in the primary semantics with the truth value of the sentence "$\overline{S}$ is true in the primary semantics" in the final semantics. The semantic mechanism here is the same as with the predicate "is a diesel motor", which connects the engine type x with the truth value of the statement "x is a diesel motor". However, since in the final semantics the truth predicate connects the truth values of two semantics, and not engines and truth values, confusion can easily occur if we don't take care which truth value belongs to which semantics. Just as the predicate "is a diesel motor" is not a logical symbol of language, because we must examine the external reality of language—a given engine—to determine the truth of the corresponding sentence, so too, the truth predicate "is true in the primary semantics" is not a logical symbol of the metalanguage (the final semantics) because we have to investigate the external reality of the metalanguage—investigate the truth value of a sentence in the original language (the primary semantics)—to determine the truth value of that sentence in the final semantics. The semantic shift that allowed us to complete the analysis of the Liar paradox, and that allows us to complete the analysis of other paradoxes of truth, as will be shown later, is precisely this change of the role of the truth predicate of the primary semantics from the logical symbol of the primary semantics, which connects the truth values within the primary semantics, into a non-logical symbol of the final semantics, which connects the truth values of the primary and the final semantics. This change leads to a change in the overall semantics of the language—from the partial two-valued primary semantics to the classical two-valued final semantics of the language.

It is easy to legalise this intuition about semantic shift. Sentences of a language with the truth predicate will always have the same meaning, but the language will have two valuation schemes—the primary and the final truth valuation. In both semantics the meaning of the truth predicate is the same: $\mathrm{T}(\overline{\varphi})$ means that φ is true in the primary semantics. But the valuation of the truth value of the atomic sentence $\mathrm{T}(\overline{\varphi})$ is different. While in the primary semantics the truth value conditions for $\mathrm{T}(\overline{\varphi})$ are classical (the truth of $\mathrm{T}(\overline{\varphi})$ means the truth of φ, the falsehood of $\mathrm{T}(\overline{\varphi})$ means the falsehood of φ, and consequently $\mathrm{T}(\overline{\varphi})$ is undetermined just when φ is undetermined), in the final semantics it is not so. In it, the truth of $\mathrm{T}(\overline{\varphi})$ means that φ is true in the primary semantics, and falsehood of $\mathrm{T}(\overline{\varphi})$ means that φ is not true in the primary semantics. It does not mean that it is false in the primary semantics, but that it is false or undetermined. So, formally looking, in the final semantics $\mathrm{T}(\overline{\varphi})$ inherits truth from the primary semantics, while other values transform to falsehood. That is why we say that this semantics is the *classical semantic closure* of the primary semantics, or in full terminology, the classical semantic closure of LIFPSK3. Due to the monotonicity of the primary semantics this means that the final semantics supplements the primary semantics in the area of its silence. If a sentence in the primary semantics has a classical value (*True* or *False*), it will have that value in the final semantics as well. If a sentence is undetermined in the primary semantics (a paradoxical sentence) then it will have a classical truth value in the final semantics that just carries information about its indeterminacy in the primary semantics. Therefore, the final semantics is the classical two-valued semantics of the language that has for its subject precisely the primary semantics of the language, and it extends the primary semantics in the part where the primary semantics is silent, using just the information about the silence.

We can see best that this is a right and a complete description of the valuation in the primary semantics by introducing predicates for other truth values in the primary valuation:

$\mathrm{F}(\overline{\phi})$ ("ϕ is false in the primary semantics") ↔ $\mathrm{T}(\overline{\textit{not}\ \phi})$
$\mathrm{U}(\overline{\phi})$("$\varphi$ is undetermined in the primary semantics") ↔*not* $\mathrm{T}(\overline{\varphi})$ *and not* $\mathrm{F}(\overline{\varphi})$

According to the truth value of the sentence φ in the primary semantics we determine which of the previous sentences are true and which are false in the final semantics. For example, if φ is false in the primary semantics then $\mathtt{F}(\overline{\varphi})$ is true while others ($\mathtt{T}(\overline{\varphi})$ and $\mathtt{U}(\overline{\varphi})$) are false in the final semantics. Once the final two-valued valuations of atomic sentences are determined in this way, the final valuation of every sentence is determined by means of the classical truth value conditions and the principle of recursion. This valuation not only preserves the primary logical meaning of the truth predicate (as the truth predicate of the primary semantics) but it also coincides with the primary valuation where it is determined.

I think that when a language user is confronted with a paradox of truth, his thinking ends in this final semantics. Therefore, the solution to the paradoxes of truth should include this semantics. Although both the primary and the final semantics share the same linguistic forms, the final semantics is the minimum metalanguage for the primary semantics by which we complete the analysis of paradoxical situations.

In the end of his article, Kripke warns that the complete description of paradoxical situations in a language with the truth predicate belongs to a metalanguage which has its own concept of truth, so the analysis of the concept of truth with fixed points remains incomplete, as in Tarski's approach. Kripke writes (Kripke 1975: 714):

> The necessity to ascend to a metalanguage may be one of the weaknesses of the present theory. The ghost of the Tarski hierarchy is still with us.

I do not think that the existence of a metalanguage with its concept of truth means that the analysis conducted here is incomplete. As already discussed, such a view arises from mixing various aspects of the concept of truth. The aim of this analysis is the logical concept of truth described on the page 11. It differs from the aspect of the concept of truth that is most important to us—truth that discriminates what is and what is not in the world that a language speaks of. The latter aspect of truth belongs to the external semantics of the language, its connection with the world, while this logical aspect of the concept of truth belongs to the internal semantics of the language. The critique of resorting to a metalanguage cannot be applied to the logical concept of truth because the truth values we associate with sentences of the metalanguage do not fall under the logical concept of truth. In particular, the concept of truth in the final semantics is not a logical concept of truth. It is equal to the concept of truth in other sciences. Of course, as in the languages of mechanical engineering, the question of the truth of sentences in the final semantics can be discussed in an appropriate metalanguage (and I've been doing it all along in these considerations). But this is a different type of problem than the problem of paradoxical sentences.[17]

[17] This is a problem of the truth regress: whenever we express a statement, we express its truth value with another statement whose truth value we express with another statement, etc.

4. *Conquering the Liar*

Having in mind this double semantics of the language (triple, if we also count the classical naive semantics), we can easily solve all truth paradoxes. On an intuitive level we have already done it for the Liar:

L: $\mathrm{F}(\overline{L})$ ("This sentence is false.")

The form of the solution is always the same. A paradox in classical thinking means that the truth value of a sentence is undetermined in the primary semantics. But then it becomes an information in the final semantics with which we can conclude the truth value of the sentence in the final semantics. To make it easier to track solutions to other paradoxes, I will sometimes distinguish by appropriate prefixes what the truth valuation is about: I will put prefix "p" for the primary semantics and prefix "f" for the final semantics. In that way we will distinguish for example "f-falsehood" and "p-falsehood".

The Strengthened Liar is "the revenge of the Liar" for solutions that seek a way out in truth value gaps, i.e., in the introduction of the third value—*Undetermined*:

SL: *not* $\mathrm{T}(\overline{SL})$ ("This sentence is not true.")

In the classical semantics it leads to a contradiction in the same way as the Liar because there "not to be true" is the same as "to be false". The paradox is used as an argument against the third value in the following way (e.g., in (Burge 1979)). If we accept that The Strengthened Liar takes on the value *Undetermined*, it means that what it is saying is true—that it is not true (but undetermined)—and so the contradiction is renewed. However, the last step is wrong because a semantic shift has occurred. The conclusion that The Strengthened Liar is undetermined is the conclusion in the final semantics. So, when we say in the end that what he says is true, this is the concept of truth of the final semantics, while the concept of truth The Strengthened Liar mentions is the concept of truth of the primary semantics. So, the truth of the final semantics is that The Strengthened Liar is not true in the primary semantics.

It is interesting that the whole argumentation can be done directly in the final semantics, not indirectly by stating the failure of the classical procedure. The argumentation is the following. If *SL* were f-false, then it would be f-false what it said—that it is not p-true. So, it would be p-true. But it means (because the final semantics extends the primary one) that it would be f-true and it is a contradiction with the assumption. So, *SL* is f-true. This statement does not lead to a contradiction but to an additional information. Namely, it follows that what *SL* talks about is f-true—that it is not p-true. So, it is p-false or p-undetermined. If it were p-false it would be f-false too, and this is a contradiction. So, it is p-undetermined.

Note that, although the Liar and the Strengthened Liar are both p-undetermined, the latter is f-true while the former is f-false.

In (Burge 1979), Burge introduces the following the revenge of the Liar for the truth value gaps solutions:

BL: $\mathtt{F}(\overline{BL})$ or $\mathtt{U}(\overline{BL})$ ("This sentence is false or undetermined.")

When we consider it in the classical semantics, if it were true then it would be false or undetermined, which is a contradiction. If it were false, then it would be true—again a contradiction. So, again we make a semantic shift and in the final semantics we conclude that it is undetermined. This means that in the final semantics it is true. Or, if we express ourselves with prefixes, that sentence is p-undetermined and f-true.

The semantic shift in argumentation is best seen in the following variant, the so-called Metaliar:

1. The sentence on line 1 is not true.
2. The sentence on line 1 is not true.

The sentence on line 1 is The Strengthened Liar so it is undetermined. If we understand the second sentence as reflection on the first sentence, which we have determined to be undetermined, then the second sentence is true. So, it turns out that one and the same sentence is both undetermined and true. In (Gaifman 1992), Gaifman uses this example to motivate the association of truth values not with sentences as sentence types but with sentences as sentence tokens. Thus, Gaifman solves the paradox by separating the same sentence type into two tokens of which the first is undetermined and the second true. In my approach, it is precisely the separation of the primary and the final semantics of the same sentence. In the 1st line it gets the undetermined value in the primary semantics, while in the 2nd, by reflection on the primary semantics, it gets the value *True* in the final semantics.

In (Skyrms 1984), Skyrms introduced the Intensional Liar, to point out the intensional character of the Liar. Namely, if in The Strengthened Liar

(1): (1) is not true.

we replace (1) with the standard name of the sentence denoted by that sign, we get the sentence

"(1) is not true" is not true.

While sentence (1) is undetermined, this harmless substitution seems to have given us the sentence which is not undetermined but true (because "(1) is not true" is undetermined, and so it is not true). But here, too, there has been a semantic shift in the truth valuation that we can clarify with prefixes:

" "(1) is not true" is not p-true" is f-true.

5. *Conquering the companions of the Liar*

In the same way, paradoxes that have a different type of failure of the classical procedure, such as the Yablo paradox (Yablo 1993), are solved. Consider the following infinite set of sentences (i), $i \in \mathrm{N}$:

(i) For all $k > i$ (k) is not true.

If the sentence (i) were true, then all the following sentences would not be true. But that would mean on the one hand that $(i+1)$ is not true, and on the other hand, since all the sentences after it are not true, that $(i+1)$ is true. So, all the above sentences are not true. But if we look what they claim entails that they are all true. This contradiction in the classical semantics turns into a true claim of the final semantics that all these sentences are p-undetermined. From what they say about their primary semantics, as with the Strengthened Liar, it follows that they are all f-true.

That the solution of the problem of the paradoxes of truth presented here is not related to negation will be illustrated by the example of Curry's paradox (Curry 1942):

C: $\mathrm{T}(\overline{C}) \rightarrow l$ ("If this sentence is true then l")

where l is any false statement. On the intuitive level, if C were false then its antecedent $\mathrm{T}(\overline{C})$ is false, and so the whole conditional C is true: we got a contradiction. If C was true then the whole conditional (C) and its antecedent $\mathrm{T}(\overline{C})$ would be true, and so the consequent l would be true, which is impossible with the choice of l as a false sentence. Therefore, we conclude in the final semantics that C is p-undetermined, and so it is f-true (because the antecedent is f-false).

All the paradoxical sentences analysed above led to contradictions in the classical semantics. Thus, in the final semantics, we concluded that they are undetermined in the primary semantics, from which we further determined their truth value in the final semantics. We could also analyse them directly in the final semantics, as was done with the Strengthened Liar. There, the contradiction would turn into a positive classical two-valued argumentation by which we would determine its truth value in both the primary and the final semantics. However, the situation is different with paradoxes which do not lead to a contradiction, which permit more valuations, like the Truth-teller. The analyses of the Truth-teller gives that it is p-undetermined. It implies that it is not p-true which means that (I: $\mathrm{T}(\overline{I})$) it is not f-true. So, I is f-false. However, although the conclusion is formulated in the final semantics, the reasoning that led to that conclusion cannot be formulated in the final semantics because it involves the analysis of the corresponding semantic graph. Of course, if we enrich the metalanguage with the description of semantic graphs and their truth value valuations then we could translate the whole intuitive argumentation into the final semantics.

In (Gupta 1982), Gupta gave several arguments against Kripke's fixed points. The solution presented here includes LIFPSK3, so this critique also applies to the solution developed in this article.

One of Gupta's criticisms, which has already been present in the literature, is that not all classical laws of logic are valid in fixed points. E.g., for a language containing the Liar, the logical law ∀*x not* (T(*x*) *and not* T(*x*)) is undetermined in each fixed point of the SK3 semantics (if we choose the Liar for *x*, we get the undetermined sentence). But since the analysis of paradoxes cannot avoid the presence of sentences that have no classical truth value, the analysis naturally leads to a three-valued language for which we cannot expect the logical laws of a two-valued language to apply. However, the SK3 semantics is maximally adapted to the two-valued logic: the logical truths of the two-valued logic are always true in SK3 when they are determined. Furthermore, the transition to the final semantics definitely solves this problem because that semantics is two-valued, and ∀*x not* (T(*x*) *and not* T(*x*)) is true in this semantics.

A somewhat more inconvenient situation is that ∀*x not* (T(*x*) *and not* T(*x*)), like other logical laws, is not true in the minimal fixed point even when there is not the Liar like or the Truth-teller like sentences. Namely, then the stated logical law is not true for its own sake—to determine its truth, the truth of all sentences, including itself, must be examined. In this way it can be seen that it is an ungrounded sentence, i.e., undetermined in MIFPSK3. But in LIFPSK3, it is true. We can easily check this by trying to give it a classic truth value. Namely, to examine its truth, we must examine whether the condition *not* (T(*x*) *and not* T(*x*)) is valid for each sentence *x*. Since we assume that language has no paradoxical sentences, it is only necessary to examine whether this is true of the law itself. If the law is false, then this condition is true of the law, so the law itself is true: we get a contradiction. Thus, the law must be true, and it is easy to show that this truth value does not lead to contradiction. Since the procedure of determining a truth value has assigned a unique truth value to this logical law, it is true in LIFPSK3. It means that this Gupta's critique turns into an argument for LIFPSK3.

The second type of Gupta's critique seeks to show that some quite intuitive considerations about the concept of truth are inconsistent with the fixed points of SK3 semantics. Gupta constructed the following example in (Gupta 1982) (Gupta's paradox). Let us have the following statements of persons *A* i *B*:

A says:

(a1) Two plus two is three. (false)
(a2) Snow is always black. (false)
(a3) Everything *B* says is true. ( )
(a4) Ten is a prime number. (false)
(a5) Something *B* says is not true. ( )

B says:

(b1) One plus one is two. (true)
(b2) My name is *B*. (true)
(b3) Snow is sometimes white. (true)
(b4) At most one thing *A* says is true. ( )

Sentences (a1), (a2), (a4), (b1), (b2) and (b3) are determined in each fixed point. However, (a3) and (a5) "wait" (b4), and (b4) "waits" them and so those sentences remain undetermined in the minimal fixed point. But on an intuitive level, it is quite easy for them to determine the classical truth value. Since (a3) and (a5) are contradictory, and all other statements of *A* are false, (b4) is true. But this means that (a3) is true and (a5) is false. However, this intuition coincides with the truth valuation in LIFPSK3. Thus, this Gupta's critique also turns into an argument for LIFPSK3. To find an intuitive counterexample for LIFPSK3 as well, Gupta replaces (a3) and (a5) with the following statements:

(a3*): (a3*) is true. ( )
(a5*): "(a3*) is not true" is true. ( )

Now at LIFPSK3, (a3*) and (a5*), and thus (b4), are undetermined. Gupta considers that on an intuitive level (b4) is true, because at most one of (a3*) and (a5*) is true. But in this step Gupta made a semantic shift from the primary semantics to the final, so (b4) is a true statement in the final semantics. This devalues his argument against LIPSK3.

6. *An erroneous critique of Kripke-Feferman theory*

In this last section I would like to draw attention to one erroneous critique of Kripke-Feferman axiomatic theory of truth (KF) which is present in contemporary literature, for example, in two contemporary respectable books on formal theories of truth. The models of this theory are the classical semantic closures of the fixed points of the SK3 semantics, and so the final semantics described in this paper, too.

In (Horsten 2011: 127) is the following text:

> So far, it seems that KF is an attractive theory of truth. However, we now turn to properties of KF that disqualify it from ever becoming our favourite theory of truth.
> Corollary 70: KF $\vdash L \wedge \neg \mathrm{T}(L)$, where L is the [strengthened] liar sentence.[18]
> …
> In other words, KF proves sentences that by its own lights are untrue. This does not look good. To prove sentences that by one's own lights are untrue seems a sure mark of philosophical unsoundness: It seems that KF falls prey to the strengthened liar problem.

[18] $\wedge$ and $\neg$ are the standard symbols for *and* and *not*.

In (Beall et al. 2018: 76) is the following text:

> But on the properties of truth itself, KF also has some features some have found undesirable. One example (discussed at length in Horsten 2011) is that KF $\vdash \lambda \wedge \neg T\lambda$. Unlike FS, KF gives us a verdict on Liars. But it seems to then deny its own accuracy, as it first proves λ, and then denies its truth. This makes the truth predicate of KF awkward in some important ways.

Both quoted texts repeat KF's critique dating back to Reinhardt (Reinhardt 1986), that axiomatic KF theory without additional restrictions is not an acceptable theory of truth. This means that its models, the classical semantic closures of the fixed points of SK3, are not acceptable solutions to the concept of truth. The reason is that the theory proves both The Strengthened Liar and that The Strengthened Liar is not true. The error in this reasoning stems from the indistinguishability of the primary (fixed point) and the final (classical semantic closure of the fixed point) semantics. KF has the role of axiomatically organising the final semantics, and what KF deduces are the true statements of the final semantics about the truth values of the primary semantics. We have already seen that The Strengthened Liar *SL* is true in the final semantics. Since KF axioms are valid in the final semantics, that KF $\vdash SL$ is not awkward but testifies to the strength of KF in the description of the fixed points. Furthermore, since *SL* is true in the final semantics, it means that it is not true in the primary semantics. So, that KF $\vdash \neg T(\overline{SL})$ is also not awkward but testifies to the strength of KF. These claims (in fact one claim KF $\vdash SL \wedge \neg T\left(\overline{SL}\right)$) are not contradictory, because different concepts of truth are involved.

References

Beall, J. 2016. "Off-topic: A New Interpretation of Weak-Kleene Logic." *Australasian Journal of Logic* 13 (6).

Beall, J., Glanzberg, M., and Ripley, D. 2018. *Formal Theories of Truth*. Oxford: Oxford University Press.

Beall, J., Glanzberg, M., and Ripley, D. 2020. "Liar Paradox." In E. N. Zalta (ed.). *The Stanford Encyclopaedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Burge, T. 1979. "Semantical Paradox." *Journal of Philosophy* 76: 169–198.

Chihara, C. 1979. "The Semantic Paradoxes: A Diagnostic Investigation." *Philosophical Review* 88 (4): 590–618.

Čulina, B. 2001. "The Concept of Truth". *Synthese* 126: 339–360.

Čulina, B. 2004. *Modelling the Concept of Truth Using the Largest Intrinsic Fixed Point of the Strong Kleene Three Valued Semantics* (*in Croatian Language*). PhD thesis. https://philpapers.org/archive/CULMTC.pdf.

Čulina, B. 2020. "The Synthetic Concept of Truth." Unpublished.

Čulina, B. 2021a. "The Language Essence of Rational Cognition with Some Philosophical Consequences". *Tesis* (*Lima*), 14 (19): 631–656.

Čulina, B. 2021b. "What is Logical in First Order Logic." Unpublished.

Curry, H. B. 1942. "The Inconsistency of Certain Formal Logics". *Journal of Symbolic Logic*, 7: 115–117.

Gaifman, H. 1992. "Pointers to Truth." *Journal of Philosophy* 89: 223–261.

Glanzberg, M. (ed.) 2018. *The Oxford Handbook of Truth.* Oxford: Oxford University Press.

Gupta, A. 1982. "Truth and Paradox." *Journal of Philosophical Logic* 11: 1–60.

Horsten, L. 2011. *The Tarskian Turn. Deflationism and Axiomatic Truth.* Cambridge: MIT Press.

Horsten, L. 2015. "One Hundred Years of Semantic Paradox." *Journal of Philosophical Logic* 44: 681–695.

Kaye, R. 1991. *Models of Peano Arithmetic.* Oxford: Clarendon Press.

Kremer, M. 1988. "Kripke and the Logic of Truth." *Journal of Philosophical Logic* 17 (3): 225–278.

Kripke, S. A. 1975. "Outline of a Theory of Truth." *Journal of Philosophy* 72: 690–716.

Quine, W. V. 1986. *Philosophy of Logic: Second Edition.* Cambridge: Harvard University Press.

Reinhardt, W. N. 1986. "Some Remarks on Extending and Interpreting Theories with a Partial Predicate for Truth." *Journal of Philosophical Logic* 15 (2): 219–251.

Skyrms, B. 1984. "Intensional Aspects of Semantical Self-Reference". In R. L. Martin (ed.). *Recent Essays on Truth and the Liar Paradox.* Oxford: Oxford University Press.

Tarski, A. 1933. "Pojęcie prawdy w językach nauk dedukcyjnych". *Towarzystwo Naukowe Warszawskie.* German translation, "Der Wahrheitsbegriff in den formalisierten Sprachen", Studia philosophica 1, 1935, 261–405.

Tarski, A. 1944. "The Semantic Conception of Truth". *Philosophy and Phenomenological Research* 4: 341–376.

Tarski, A. 1969. "Truth and Proof". *Scientific American* 220 (6): 63–77.

Visser, A. 1989. "Semantics and the Liar Paradox". In D. M. Gabbay et al. (eds.). *Handbook of Philosophical Logic.* Volume 4. Reidel: Springer.

Yablo, S. 1982. "Grounding, Dependence, and Paradox." *Journal of Philosophical Logic* 11 (1): 117–137.

Yablo, S. 1993. "Paradox without Self-Reference." *Analysis* 53: 251–252.

Croatian Journal of Philosophy
Vol. XXIII, No. 67, 2023
https://doi.org/10.52685/cjp.23.67.2
Received: May 16, 2022
Accepted: January 7, 2023

Evolutionary Game Theory and Interdisciplinary Integration

WALTER VEIT*
University of Bristol, Bristol, UK
Ludwig-Maximilians-Universität München, Germany

Interdisciplinary research is becoming more and more popular. Many funding bodies encourage interdisciplinarity, as a criterion that promises scientific progress. Traditionally this has been linked to the idea of integrating or unifying disciplines. Using evolutionary game theory as a case study, Till Grüne-Yanoff (2016) argued that there is no such necessary link between interdisciplinary success and integration. Contrary to this, this paper argues that evolutionary game theory is a genuine case of successful integration between economics and biology, shedding lights on the many dimensions along which integration can take place.

Keywords: Interdisciplinarity; integration; evolutionary game theory; biology; economics.

1. *Introduction*

For much of the 20th century, reductionism was the dominant approach in philosophy of science (see Nagel 1935, 1949, 1979). However, with the demise of logical empiricism, reductionism as a regulative ideal of science has become more and more criticized by historians and philosophers of science (see Feyerabend 1962; Kuhn 1962; Schaffner 1967). Many subfields within philosophy of science such as biology, have even developed an anti-reductionist consensus (see Kitcher 1984, 1990; Rosenberg 1985, 1994; Dupré 1993). Similar debates currently unfold

* I would like to thank the audience at Bristol who attended a talk on this paper for helpful comments. This paper is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 101018533).

in the philosophy of economics (see Sugden 2001; Fumagalli 2013). In fact, reductionism has become almost a dirty word, with only a minority willing to embrace the term as a badge of honour (see Rosenberg 2006).

Over time, reductionism has been replaced by a new ideal, i.e. unification or integration (see Kitcher 1999). According to Till Grüne-Yanoff (2016) the increasing popularity of interdisciplinary research, as a scientific virtue, is due to interdisciplinary success being linked to integration between fields or disciplines. In fact, Holbrook argues that the "notion of 'integration' is so widespread in the [interdisciplinarity] literature that to question whether [interdisciplinarity] involves integration is almost heretical" (2013: 13). However, Grüne-Yanoff (2016) argues that there is no such necessary link between interdisciplinary success and integration, contrary to what others have argued before him (Lattuca 2001; Klein 2010; Holbrook 2013).

Grüne-Yanoff illustrates his case with two separate case-studies for interdisciplinary model exchange. First, evolutionary game theory as an example of interdisciplinary exchange between economics and biology, and secondly hyperbolic discounting as an example of interdisciplinary exchange between economics and psychology. Considering the wide recognition of both examples as interdisciplinary successes, Grüne-Yanoff (2016) was wise to choose them in order to ward off objections that his case-studies do not warrant the judgement that despite interdisciplinary success there was no "integration of disciplines, concepts or methods" (2016: 344).

However, this naturally leaves him open for the opposite criticism that I spell out in this paper. Highly abstract and simplified models are, of course, used across scientific disciplines (Veit 2019a). Both economists and philosophers wary of the common criticism directed against economic models being unrealistic or unreliable have drawn on modelling practice in biology to justify and improve 'unrealistic' economic models (see Sugden 2001, 2009, 2011; Rosenberg 2009; Odenbaugh and Alexandrova 2011). In one very fascinating case, however, economists went so far as to import a model framework from biology in its entirety, i.e. evolutionary game theory, a model framework that has previously been adopted by biologists applying game-theoretic tools from economics to biology. In this paper, I argue that evolutionary game theory, contrary to Grüne-Yanoff (2016) is in fact, a case of both interdisciplinary success and integration. Nevertheless, I agree with Grüne-Yanoff's (2016) general sentiment that there is no necessary link between interdisciplinary success and integration, though their relation is stronger than he suggests.

This paper is structured as follows: Section 2 discusses how Grüne-Yanoff defines the conditions for interdisciplinary success and integration. Section 3 sketches the history of evolutionary game theory and explains the two most fundamental concepts used within it. Section 4 provides an argument that EGT has led to a methodological integration between biology and economics. Section 5 provides an argument that there has also been conceptual integration. Section 6, finally concludes the discussion.

2. *Interdisciplinary Success without Integration*

In order to understand whether EGT is a case of interdisciplinary success and integration between biology and economics, we will first require to clear up the meaning of both terms. In doing so, I closely follow Grüne-Yanoff's (2016) definitions as my disagreement merely lies in his mischaracterization of EGT. As Grüne-Yanoff (2016) points out, the first relevant question to ask is what interdisciplinary success entails and why it is valued, with a further distinction opening up by asking whether interdisciplinarity is valued as a goal in itself or only instrumentally. Grüne-Yanoff (2016: 345) cites the Economic and Social Research Council (ESRC) that justifies its funding of interdisciplinary projects by highlighting the instrumental goals that can be achieved in such a way:

> many of the most pressing research challenges are interdisciplinary in nature, both within the social sciences and between the social sciences and other areas of research. (ESRC 2013)

On the other side, Grüne-Yanoff (2016: 345) cites the director of the National Health Institute (NIH), Elias A. Zerhouni, who explains the aims of its funding projects with the goal:

> to encourage and enable change in academic research culture to make interdisciplinary research easier to conduct for scientists who wish to collaborate in unconventional ways. (NIH News 2007)

Grüne-Yanoff suggests that this is a case of interdisciplinarity valued for its own sake. However, this conclusion is far from obvious. The growing support for interdisciplinary may simply rest on the belief that *unconventional research* has historically shown to have the best prospects for achieving scientific progress, such as "detailed explanations, more accurate predictions or more effective control", examples Grüne-Yanoff lists himself (2016: 345). Evidence for this can easily be found. Some famous examples are Gregor Mendel's study of peas and Galileo's experiments that were at least at the time unusual. If so, interdisciplinarity would only be valued because it is unconventional and perhaps requires financial support to bridge the gaps between disciplines. Hence, there need not be a necessary link between interdisciplinarity and the goal of unification, a conclusion Grüne-Yanoff would certainly embrace, but it is not so clear that the view he is attacking is actually in the majority.

Nevertheless, Grüne-Yanoff (2016) provides a useful and succinct philosophical analysis of the literature on interdisciplinarity with two criteria emerging on which interdisciplinarity can be understood. First: "the disciplines involved in interdisciplinary interaction change their identity in some relevant way" (346). Second: "the change that disciplines undergo in successful interdisciplinary exchanges leads them to integrate in a relevant way" (346). Though Grüne-Yanoff agrees with the former, he argues against the latter. As already alluded to, I agree with this general sentiment of his argument. However, the connections between integration and interdisciplinarity are deeper than he himself suggests. The first criterion provides a straight-forward case for measuring interdisciplinary exchange (though not necessarily interdisciplinary success). In the case of evolutionary game theory, it is already widely agreed that the application of game theory to biology and the use of EGT in economics has been quite successful. Whether the disciplines changed in a relevant way is less obviously clear, and stands in an direct relationship with the degree of integration taking place.

The case Grüne-Yanoff (2016) makes is a perhaps unintuitive, but possible: disciplines can change by *de-integrating*, i.e. moving further apart. This may seem unappealing, but as Grüne-Yanoff successfully argues it is a real possibility and could nevertheless qualify as scientific progress. Grüne-Yanoff is aware of this connection and points to the unificationist ideas that underlie the arguments from "defenders of the interdisciplinary-as-integration" (2016: 348) view, such as Klein: "the roots of the concepts lie in a number of ideas that resonate through modern discourse—the ideas of a unified science, general knowledge, synthesis and the integration of knowledge" (Klein 1990: 19). As alluded to in the introduction, reductionism has become less and less popular among philosophers of science. The unity of science thesis by Carnap was untenable given its strong formulation: "science is a unity, [such] that all empirical statements can be expressed in a single language, all states of affairs are of one kind and are known by the same method" (Carnap 1934: 32). However, the disunity of science was and is a position, many philosophers of science would like to avoid, hence leading to a variety of less strict conditions for the unification of science (see Kitcher 1999; Brigandt 2010). One such alternative is integration. The key then is to understand what integration entails.

Grüne-Yanoff (2016) summarizes the literature on integration and comes to several conclusions. Firstly, integration goes beyond mere theory: it "affects the concepts they use, both in their explanations, as well as in their ontological content" and it "affects their practices, specifically their terminology, their methods and their data" (2016: 347). Secondly, integration can be measured by the increase in overlap in at least one of these categories (see O'Malley 2013; Grüne-Yanoff 2016). A strong view on the necessary link between interdisciplinary success and integration emerges that Grüne-Yanoff characterizes as follows:

The Strong View (SV):

> "[I]nterdisciplinary research is successful if it integrates disciplines, creates *new academic programs and ultimately new disciplines*." [italics added] (Grüne-Yanoff 2016: 348)

As successful *de-integration* of disciplines is a real possibility, the SV is literally too strong. In fact, some authors such as van der Steen (1993) have explicitly argued for the de-integration of scientific fields, such as biology, due to the danger of overgeneralization. One example within biology is the use of different notions of functions (see Garson 2017) and genes (see Rosenberg 2006) within different sub-fields. Historically, much confusion has been created by authors who interpreted terms differently, for example during the group selection debate (see Okasha 2006, Veit 2019b). So even though biology might seem like a field, where integration seems to be an inherently valuable goal, it could come at a severe cost if unification is merely searched for the sake of unification. Furthermore, as Grüne-Yanoff points out: the failure of an attempt to integrate may simply be explained by the fact that the two disciplines cannot be unified (see O'Malley 2013). Hence, contrary to Klein (2008), failing in an attempt to integrate two disciplines need not imply failure.

Nevertheless, the SV highlights a possibility Grüne-Yanoff has disregarded, i.e. *the emergence of new academic programs and disciplines*. In the following, I am going to argue that despite differences in microstructure between biology and economics, EGT has developed a sophisticated set of models to deal with a macro-phenomena common to both. Unrecognized by Grüne-Yanoff, this has led to the creation of a new field, i.e. the field of *evolutionary dynamics*. In the following section, I characterize the history of EGT and point out some differences to Grüne-Yanoff's analysis offered across multiple papers (2011a, 2011b, 2013, 2016).

3. *The history of evolutionary game theory*

Now widely used in biology and the social sciences, though primarily economics, EGT has had an interesting history of success. In the following, I provide a short history and explain the development of the most important tools of EGT: The equilibrium concept of an *evolutionary stable strategy* (ESS), introduced by Maynard Smith and Price (1973) and the formal equation of the *replicator dynamics* introduced by Taylor and Jonker (1978). Though these tools are used in both biology and the social sciences and share the same formal framework and equations, they often have to be interpreted differently depending on the discipline.

EGT is most often associated with John Maynard Smith, who together with George Price (1973) introduced the concept of an ESS to analyse conflicts between animals. More broadly they introduced EGT to explore questions regarding how well a phenotype does, depending on the phenotypes present in a population, i.e. frequency-dependent fitness. The first traces of such a methodology, however, can be traced back as far as 1930, when R. A. Fisher (1930), worked on a mathematical solution to explain the equal sex ratio in animals. As a vast number of field studies shows, the majority of males in many species do not reproduce suggesting the benefit of a female-biased sex ratio. Fisher argued that the equal sex ratio can be explained by treating this situation as a game of strategic interaction. If the population consists of a majority of females, male offspring will have a higher expected fitness value than female offspring until their share in the population evens out, despite the fact that the actual fitness of many males will be zero. As this example shows, strategies are a central component in evolution and it was only natural that game theory could be successfully applied to biology (Veit 2021a).

According to Maynard Smith previous models of evolution have been insufficient to analyse three common characteristics: “group selection, kin selection and frequency-dependent selection” (1974: 210). What EGT provides, is a formalism in which all of these explanatory strategies can be captured in the terms of individuals, their strategies and associated fitness. Surprisingly, Maynard Smith himself initially took this formalism to be almost so simplistic that it could only be trivial.

Nowadays, however, EGT has illuminated many problems such as the evolution of cooperation, trust and language (Veit 2019c). Given its origin, the structure of EGT, naturally, bears great resemblance to the individualism espoused in game theoretic explanations of social behaviour, with individuals, their strategies and preferences over outcomes, i.e. utility. Game theory was invented thirty years prior by von Neumann and Morgenstern in the *Theory of Games and Economic Behavior* (1944) and has become one of the most influential works in the social sciences. During a stay at the University of Chicago, Maynard Smith was so enamoured with the simplicity and generality of game theoretic tools that he was led to adopt the formal structure of game theory for problems in biology. However, seemingly supporting Grüne-Yanoff's argument, Maynard Smith did not think of his work as an integration between biology and economics, something that is emphasized by the following quote from Maynard Smith's influential book *Evolution and the Theory of Games*:

> Sensibly enough, a central assumption of classical game theory is that the players will behave rationally, and according to some criterion of self-interest. Such an assumption would clearly be out of place in an evolutionary context. Instead, the criterion of rationality is replaced by that of population dynamics and stability, and the criterion of self-interest by Darwinian fitness. (Maynard Smith 1982: 2)

Since then, models and simulations have become an integral part of the biologist's toolkit. Back when Maynard Smith introduced EGT, however, many biologists where openly hostile to the mathematization of the discipline. In fact, the *Journal of Theoretical Biology*, in which Maynard published a more extensive treatment of his idea to import game theory into biology (1974) was only founded in 1961. Maynard Smith, who served as an engineer for civil planes during the second world war, was familiar with the use of highly idealized models, in fact, he knew that one could put faith in them even when human lives where at stake. "I also acquired the ability, rare among biologists, to perform massive numerical operations [...] and without making mistakes; a mistake could mean that someone got killed" (1985: 349). His trust in the power of mathematical models would later lead him to study under J. B. S. Haldane and apply his acquired modelling skills to biological problems.

In game theory, institutions and social phenomena are fully accounted for in terms of individual choices. This underlying individualism is also the methodology of EGT. Instead of a kind of biological holism accounting for its complexity, EGT espouses a mechanistic form of empirical research. Maynard Smith, rather than advocating the use of dubious concepts such the *good of the species*[1] aimed to explain apparently unfit behaviour, such as altruistic warning calls that alert the group of a predator, but putting the individuals own fitness at risk, purely in terms of kin-selection. As we shall see EGT models are often directed against impossibility[2] claims according to which selection on the level of the individual could not be responsible for the evolution of cooperation and altruism. In EGT, underlying mechanisms such as kin-selection or frequency-dependent selection are to be analysed isolated from interfering forces. Naturally, this takes away much of the realism from the model world that is created with only loose resemblance to the real world. However, Maynard Smith (1974) argues that it is necessary to start from very simple assumptions to learn about the mechanism itself. Whether the hypothesized mechanism operates in the real world is a distinct, albeit important question. Cognitively limited agents such as us could otherwise not understand complex phenomena in economics and biology. This abstraction is, as I shall argue in the next section, the key towards understanding how EGT integrated biology and economics.

However, let us first take a look at the process of building an EGT model. Unlike game theory, EGT models do not maximize utility but fitness, i.e. reproductive success. While it might be impossible to unify human desires into a single utility scale, the concept of fitness allows for a comparatively straightforward way of assigning values to outcomes. For players to rank their preferences and make coherent choices, game theory assumes players to be rational. EGT, on the other hand, does not even require the 'players' to be conscious. Strategies are hard-wired behaviour, or more broadly, alternative phenotypes. Unlike rational agents, individuals in EGT can truly *just be* animals unaware of the game they are playing. Not even the ability to 'play' a different strategy is a necessary requirement, as long as strategies are passed on to one's offspring. While there are many refinements of the Nash equilibrium

[1] A thesis endorsed by influential biologists such as Wynne-Edwards (1962) and Konrad Lorenz (1966).

[2] Or at least near impossibility.

in game theory, each liable to criticism, EGT employs multiple stability solution concepts: the most famous one being the evolutionary stable strategy (ESS) provided in John Maynard Smith and Price (1973). If a strategy i is evolutionary stable, there cannot be another invading strategy j with a higher fitness, i.e. $u(i) > u(j)$. Hence, the payoff u of a member of the population playing i against another member playing i must be higher than a mutant playing j against a member of the population playing i, or if their payoff is the same, the incumbent strategy must do better against a mutant than the mutant would do playing against another mutant. The interaction payoffs can be represented formally as follows:

$$u(i,i) > u(j,i) \tag{1.1}$$

Or

$$u(i,i) = u(j,i) \quad and \quad u(i,j) > u(j,j) \tag{1.2}$$

The ESS captures a Nash equilibrium (NE), i.e. condition 1.1, in which, the equilibrium cannot be invaded by a low share of mutants playing an alternative strategy. Hence, every ESS is a NE but not every NE is a ESS. However, just like the possibility of multiple NE, this refinement of the NE allows for multiple ESS. In which state a population ends up depends upon the initial conditions. Let us take a look at Maynard Smith's original and most famous EGT model, the highly idealized *Hawk-Dove Game*[3]:

Table 1 *The payoff matrix for the* Hawk-Dove Game

	Hawk	Dove
Hawk	$\frac{1}{2}$ (V – C)	V
Dove	0	V/2

In their simplest form, EGT models represent two-player games within populations that are infinite, with interactions happening at random and consisting of indistinguishable individuals.[4] In the *Hawk-Dove Game*, there are only two pure strategies in response to a resource contest: Hawk refers to the aggressive strategy leading either to the withdrawal of the opponent or an escalated conflict, i.e. battle with the cost of a potential injury C. Dove refers to the passive strategy of displaying and retreating when the opponent escalates. If a Hawk meets a Dove it will always win and receive a resource associated with a value V. Both V and C are expressed in terms of change in fitness. Hence, if V > C, i.e.

[3] Based on an updated treatment in Maynard Smith (1982) *Evolution and the Theory of Games.*

[4] All of these assumptions can made more realistic leading to agent based models, e.g. finite populations or the introduction of population structure via cellular automata.

the value of the resource for reproduction is higher than the negative effect of an injury on fitness, Hawk would be the dominant strategy. Doves would be driven to extinction, even when there is only a single Hawk mutant in a Dove population. However, when C > V the result will be a mixed strategy. Even though Hawks always win against Doves, they risk injury when meeting other Hawks. Doves encountering other Doves, on the other hand, share the resource. Whereas mixed strategies in game theory are randomizations, in EGT mixed, ESS are either stable polymorphic populations playing pure strategies or randomized but encoded strategies in individuals. The mixed strategy can be calculated by solving the following equation:

$$u(H,I) = u(D,I) \tag{1.3}$$

The result of solving equation (1.3) is P = V/C with P representing the share of Hawks or the probability of individuals playing Hawk.[5] If the fitness value of the resource is 1 and the cost of fighting 2, or generally twice as large as the value of the resource the population will be in a mixed equilibrium with either 50% playing Hawk and 50% playing Dove or a mixed strategy randomizing between Hawk and Dove. By putting these arbitrary values into the payoff-matrix, this result can be easily illustrated:

Table 2 *The payoff matrix for the* Hawk-Dove Game*

	Hawk	Dove
Hawk	−0.5	1
Dove	0	0.5

Only when the population plays Hawk and Dove with equal probability of 50% are the payoffs for both strategies equal, i.e. an expected fitness value of 0.25. Here numbers do not refer to any real properties of the real world but rather the logical possibilities of symmetric contests within a *model population*. Such conceptual exploration of a model is familiar from economic modelling practice. In order to increase the realism of the model, this game has been extended in various ways, most importantly through the addition of asymmetric cues.

However, several authors (see Huttegger and Zollman 2012, 2013) argue that the generality and simplicity of a fundamentally static concept such as the ESS faces severe limits in understanding the dynamics of evolutionary processes. For the purposes of this paper, I can only reiterate their call for a pluralistic methodology (see also Veit 2021b), employing both static and dynamic game theoretic tools, some of which

[5] A mathematical proof for this result is provided in the very same book by Maynard Smith (1982).

originate in economics. The most famous dynamical approach in EGT goes back to Taylor and Jonker (1978), who developed the *replicator dynamics* with the explicit goal to fill the dynamical gap the ESS left. As already alluded to, EGT allows for both biological and cultural interpretations explaining the interdisciplinary interest in EGT. While the biological form of these models treats replication as inheritance, replication has to be interpreted as learning or imitation in the cultural setting. Replicator dynamics (RD) are an attempt to model the relative changes of strategies in a population. These can be either instantiated biologically or culturally. Strategies with higher fitness than the population average prosper and increase their share in the population, while those with lower fitness are driven to extinction. RD in the biological setting are thus an attempt to model the dynamics of reproduction and natural selection. The following is the continuous replicator dynamics equation:

$$(1.4) \qquad \frac{dx_i}{dt} = [u(i,x) - u(x,x)] * x_i \qquad \text{(Weibull 1995: 72)}$$

In each round individual strategies, *i* increase their share within a population linear to their success compared to the average fitness in the population. Just as the ESS, RD assume infinite population size or at least infinite divisibility and random interaction. These idealisations serve the purpose to analyse the frequency-dependent success of different strategies, whether they are biologically or culturally transmitted.

Robert Axelrod (1980) is the perhaps most famous author for applying EGT in the social sciences. Himself a political scientist, he sought to explain the emergence of cooperation. While the traditional prisoner's dilemma (PD) game from game theory seemed to suggest that defection is always the rational move, things change when games are repeated. Axelrod coined the term *tit-for-tat* as a strategy that is forgiving, starts fair and only retaliates once the opponent cheats. When the other player returns to cooperation and the tit-for-tat player notices this, he returns himself back to cooperation in the next round. When two tit-for-tat players meet, they always cooperate. Such a cooperative strategy was later observed in sticklefish (see Milinski 1987) and also given a biological interpretation by Axelrod. As Sugden (2001) and Grüne-Yanoff (2011a) point out, early economists were dissatisfied with the rationality requirements of classical game theory. Let me now turn to my criticism of Grüne-Yanoff's characterization of EGT and argue that biology and economics have indeed become more integrated. Following Grantham's (2004) distinction between theoretical and practical integration, I argue for this thesis along two lines. First, I argue that biology and economics have become integrated on a practical dimension increasing the overlap between model-building in the two disciplines. Secondly, I argue that biology and economics have become theoretically integrated, bridging the strong divide between the study of rational agents and organisms.

4. *Methodological Integration*

Compared to biology, modelers in economics rarely attempt to bridge the gap between conclusions in the model world to conclusions about the real world, even when they are using the very same formal structure for their models (see Grüne-Yannoff 2011a, 2011b). Contrary to Grüne-Yanoff (2016) I argue that despite this difference the history of EGT shows that economic and biological modelling practice, in fact, moved closer together. Perhaps due to a sort of physics envy, beginning with Robbins (1932), economists have been reluctant to use inductive methods that are widespread in biology and could have helped them to provide better explanations. Rosenberg (1992) has argued that economics rather than being a genuine scientific discipline has just been a form of applied mathematics, studying diminishing returns and optimization without any significant improvement in predictive power since Adam Smith. A significant change, however, took place when EGT was introduced into economics, something Robert Sugden calls the *evolutionary turn*:

> Evolutionary game theory is still in its infancy. A genuinely evolutionary approach to economic explanation has an enormous amount to offer; biology really is a much better role model for economics than is physics. I just hope that economists will come to see the need to emulate the empirical research methods of biology and not just its mathematical techniques. (Sugden 2001: 128)

Eight years after Sugden's article on the evolutionary turn in game theory, Rosenberg (2009) recognized the transition economics underwent in the past three decades to a discipline much closer biology, for at least three reasons: First and here agreeing with Sugden (2001), EGT provides a foundation for the results of game theory that are far less ontologically demanding. than the strong rationality requirements of classic rational choice theory. In fact, Ken Binmore in the foreword to Jörgen Weibull's book *Evolutionary Game Theory* (1995) points out that Maynard Smith led economists to reconsider their rationality assumptions that seemed to put a clear dividing line between biology and economics.

> Maynard Smith's book Evolution and the Theory of Games directed game theorists' attention away from their increasingly elaborate definitions of rationality. After all, insects can hardly be said to think at all, and so rationality cannot be so crucial if game theory somehow manages to predict their behavior under appropriate conditions. (Ken Binmore, foreword in Weibull 1995: x)

Unlike Maynard Smith criticism of economic modelling suggests, economists were positively thrilled about applying EGT to economics. Second, a revolution in experimental economics took place, importing models and data from psychology and neuroscience, familiar from the testing of EGT models in biology. Third, the weakening of assumptions concerning perfect information. Much work since then has been done on information and signaling games in both biology and the social sciences, often employing various EGT models (see Skyrms 2010; Grafen 1990). Recognizing that economic explanations like biological ones are "path-dependent, subject to historical contingencies, and in many respects, inherently unpredictable" (Sugden 2001: 113) should highlight how economic modelling practice moved closer to biological modelling. The practices integrated.

Hutteger and Zollman (2013) draw a new dividing line: one between biological game theory and game theory used in the social sciences. This may be a more useful distinction, as EGT has led to a new discipline applicable to both economics and biology, i.e. the field of evolutionary dynamics.

5. *Conceptual Integration*

Unlike the import of game theory from economics to biology, the import of EGT from biology to economics involved, at least in the beginning, only minor adjustments. Instead, Grüne-Yanoff argues, that "particularly in the early years" economists "explored the consequences of introducing existing formal concepts into the body of economic modelling" (2011a: 395). As alluded to in Section 4, economists and philosophers alike hoped that the methodological integration of economics and biology could lead to ontological integration. However, Grüne-Yanoff importantly points out that the biological interpretation of EGT is often incompatible with the social phenomena economist aim to explain. Grüne-Yanoff even goes so far to suggest that "[b]ecause economists lacked resources to provide a more fitting re-interpretation, they often engaged in analogy construction, as for example illustrated by the meme concept" (2011a, 395). However, the meme (see Dawkins 1976, Dennett 1995, Schlaile et al. Forthcoming) as a cultural analogy to the gene in biology, is not necessarily as problematic as Grüne-Yanoff suggests. After all, if there is a straightforward analogy to be found here, it seems hard to deny that at least some integration actually took place. Furthermore, it is unclear how the concept of memes is any more problematic than the concept of utility-maximization of rational agents. Nevertheless, evolutionary game theorists working on cultural evolution have made it clear that no entity such as memes need be postulated for EGT to work in a cultural setting (see Alexander 2009). However, the same may be said for the gene, left omitted in the biological interpretation of EGT models. Even in a contrafactual world where the genetic code was not yet discovered, these models would have considerable explanatory and predictive power.

While evolutionary game theory has undergone significant changes from the original game theory, Grüne-Yanoff (2011a, 2011b) rightly criticized economists for a myopic use and import of EGT models into their own discipline disregarding the different microstructure in biology. Concepts such as biological replication need to be replaces by learning or imitation mechanisms. However, going further Grüne-Yanoff (2013) quotes Mayntz (2004) to argue that there is, in fact, no common causal core between the biological and social mechanisms the RD represents. As I argue against this claim it is useful to take a look at the quote ourselves:

> Processes identified in the causal reconstruction of a particular case or a class of macrophenomena can be formulated as statements of mechanisms if their basic causal structure (e.g., a specific category of positive feedback) can also be found in other (classes of) cases. The mobilization process observed in a fund-raising campaign for a specific project can, for instance, be generalized to cover other outcomes such as collective protest or a patriotic movement inducing young men massively to enlist in a war. A particular case of technological innovation like the QWERTY keyboard may similarly be recognized as a case in which an innovation that has initially gained a small competitive advantage crowds out technological alternatives in the long run. This is already a mechanism of a certain generality, but it may be generalized further to the mechanism of "increasing returns," which does not only apply to technological innovations but has also been used in the analysis of institutional stability and change . . . "Increasing returns," of course, is a subcategory of positive feedback, an even more general mechanism that also operates in the bankruptcy of a firm caused by the erosion of trust or in the escalation of violence in clashes between police and demonstrators. (Mayntz 2004: 254, quoted in Grüne-Yanoff 2013: 86)

Grüne-Yanoff argues that the different interpretations of the replicator dynamics in biology and economics constitute an *isolation gap* and hence do not "share a common abstract causal structure" (2013: 83). However, though there is a gap in EGT often leaving out how strategies are replicated, I argue that Grüne-Yanoff's argument does not provide sufficient reason not to treat both cultural and biological evolution as more abstract Darwinian processes following the same causal mechanism. This question relates to the program of a generalized theory of evolution covering not only biological but also cultural evolution. As Godfrey-Smith (2009) argued, how strategies are replicated is not essential for the theory of natural selection. In fact, before the modern synthesis, Darwin's theory had no substantive, nor accurate theory of how phenotypes could be inherited. The gene-concept similarly was treated as *whatever is responsible for replication*. With progress in genetics and molecular genetics, we have gained much understanding of how this mechanism works. But natural selection was a well-established theory with considerable explanatory power well before that. What is established is no less than a mathematical truth, a theorem that predicts evolutionary change if certain conditions are met. This had made Karl Popper worried about the unfalsifiability of evolution (1976), only later changing his mind when such a position seemed to be a good *prima facie* reason to reject falsificationism itself (1978). Popper certainly would not have anticipated the now widespread use of his criterion among creationists. Because evolution is a substrate-neutral algorithm (see Dennett, 1995) and applies at every level of organization, we can have confidence that an abstract

Darwinian process operates within not only the biological but also the social realm. This is a big advantage evolutionary models share: the confidence that at their most fundamental level they are modelled with a well-established mechanism that does not rely on the demanding rationality assumptions of classical game theory. EGT models are able to explain the emergence and stability of local equilibria. Criticizing the highly abstract EGT models for particular mechanisms such as learning, imitating and reproduction are instantiated differently misses the point, whether or not a theoretical entity such as *memes* are postulated. These models share a common Darwinian core that is explored in the field of evolutionary dynamics.

6. *Conclusion*

In this paper, I argued that EGT is in fact, a paradigm case of integration between two disciplines. Contrary to Grüne-Yanoff (2016) I argued, that the history of EGT a case of both interdisciplinary success and integration. Though I agreed with the general message of Grüne-Yanoff's (2016) argument, that there is no necessary link between interdisciplinary success and integration, I argued against his claim that the history of EGT is one of de-integration between biology and economics.

Having provided a short history of how biologists adopted game theory and developed new concepts such as the ESS and the RD, I have argued that these events were a clear case of integration in methodology. During the last century, biology went from a discipline in which mathematical models were viewed as hostile, to a discipline in which mathematical models play a key role and at least fundamental mathematical skills have become a necessity to work in the field of theoretical biology and EGT played a not minor role in this shift.

Perhaps due to Darwin (1859), who himself regretted the lack of mathematical skills and provided his account of natural selection purely with verbal arguments led generations of biologists to hold the view that there is no need for mathematics in biology. Furthermore, economists started to be more concerned with the realism of their models seeking to conduct experiments, simulations and gather empirical data. But there has not only been methodological integration between the disciplines. The concept of strategic interaction plays a crucial role in modern biology, and the strong rationality assumptions of classical game theory have been weakened. Hence, the concepts in both fields have moved closer together. Perhaps most interestingly, a new field has emerged, i.e. the field of evolutionary dynamics, studying both cultural and biological evolution as instantiations of a more abstract causal process. While the integration between economics and biology might be considered relatively minimal, that is a very different conclusion than the denial that integration took place. But it is precisely these gradual and perhaps hard to see changes that historians and philosophers of science should pay attention to.

References

Alexander, J. M. 2009. "Evolutionary Game Theory." *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2009/entries/game-evolutionary

Axelrod, R. 1980. "More effective choice in the Prisoner's Dilemma." *Journal of Conflict Resolution* 24: 379–403.

Axelrod, R. M., and Hamilton, W. D. 1981. "The evolution of cooperation." *Science* 211: 1390–1396.

Carnap, R. 1934. *The Unity of Science.* London: Kegan Paul, Trench, Trubner, and Co.

Darwin, C. 1859. *On the Origin of Species.* John Murray, London [1964 facsimile edition, Cambridge: Harvard University Press].

Dawkins, R. 1976. *The Selfish Gene.* Oxford: Oxford University Press.

Dennett, D. C. 1995. *Darwin's Dangerous Idea: Evolution and the Meanings of Life.* New York: Simon and Schuster.

ESRC 2013. "Guidance for applicants." https://www.esrc.ac.uk/funding-and-guidance/applicants/. Accessed 10.01.2019.

Feyerabend, P. 1962. "Explanation, Reduction and Empiricism." In H. Feigl and G. Maxwell (ed.). S*cientific Explanation, Space, and Time* (Minnesota Studies in the Philosophy of Science, Volume III), Minneapolis: University of Minneapolis Press, 28–97.

Fisher, R. A. 1930. *The Genetic Theory of Natural Selection*, Oxford, Clarendon Press.

Fumagalli, R. 2013. "The futile search for true utility." *Economics and Philosophy* 29 (3): 325–347.

Garson, J. 2017. "How to Be a Function Pluralist." *The British Journal for the Philosophy of Science* 69: 1101–1122.

Godfrey-Smith, P. 2009. *Darwinian Populations and Natural Selection.* Oxford: Oxford University Press.

Grafen, A. 1990. "Biological signals as handicaps." *Journal of Theoretical Biology* 144 (4): 517–546.

Grantham, T. A. 2004. "Conceptualizing the (Dis) unity of science." *Philosophy of Science* 71 (2): 133–155.

Grüne-Yanoff, T. 2011a. "Models as products of interdisciplinary exchange: Evidence from evolutionary game theory." *Studies in History and Philosophy of Science* 42: 386–397.

Grüne-Yanoff, T. 2011b. "Evolutionary game theory, interpersonal comparisons and natural selection: a dilemma." *Biology and Philosophy* 26: 637–654.

Grüne-Yanoff, T. 2013. "Models of Mechanisms: The Case of the Replicator Dynamics." In H. K. Chao, S. T. Chen, R. Millstein (eds.). *Mechanism and Causality in Biology and Economics*. History, Philosophy and Theory of the Life Sciences, vol 3. Dordrecht: Springer.

Grüne-Yanoff, T. 2016. "Interdisciplinary success without integration." *European Journal for Philosophy of Science* 6 (3): 343–360.

Holbrook, J. B. 2013. "What is interdisciplinary communication? Reflections on the very idea of disciplinary integration." *Synthese* 190 (11): 1865–1879.

Huttegger, S. M. and Zollman, K. J. S. 2012. “Evolution, Dynamics, and Rationality: The Limits of ESS Methodology.” In K. Binmore and S. Okasha (eds.). *Evolution and Rationality: Decisions, Co-operation, and Strategic Behaviour*. Cambridge: Cambridge University Press, 67–83.

Huttegger, S. M. and Zollman, K. J. S. 2013. “Methodology in Biological Game Theory.” *The British Journal for the Philosophy of Science* 64: 637–658.

Kitcher, P. S. 1984, “1953 and all that. A tale of two sciences.” *The Philosophical Review* 43: 335–371.

Kitcher, P. S. 1999. “Unification as a regulative ideal.” *Perspectives on Science* 7 (3): 337–348.

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

Klein, J. T. 2008. “Evaluation of interdisciplinary and transdisciplinary research: a literature review.” *American Journal of Preventive Medicine* 35 (2): 116–123.

Klein, J. T. 2010. “A taxonomy of interdisciplinarity.” In R. Frodeman, J. T. Klein, and C. Mitcham (eds.). *The Oxford Handbook of Interdisciplinarity*. Oxford: Oxford University Press, 15–30

Kuhn, T. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lattuca, L. R. 2001. *Creating Interdisciplinarity: Interdisciplinary Research and teaching among College and University faculty*. Vanderbilt University Press.

Lorenz, K. Z. 1966. *On Aggression*. London: Methuen.

Maynard Smith, J. 1974. “The theory of games and the evolution of animal conflicts.” *Journal of Theoretical Biology* 47: 209–221.

Maynard Smith, J. 1982. “Evolution and the Theory of Games.” Cambridge: Cambridge University Press.

Maynard Smith, J. 1985. “In Haldane’s footsteps.” In D. A. Dewsbury (ed.). *Leaders in the Study of Animal Behavior: Autobiographical Perspectives*. Lewisburg: Bucknell University Press, 347–354

Maynard Smith, J. 1996. “The games lizards play.” *Nature* 380: 198–199.

Maynard Smith, J. and Parker, G. A. 1976. “The logic of asymmetric contests.” *Animal Behaviour* 24: 159–175.

Maynard Smith, J. and Price, G. 1973. “The logic of animal conflicts.” *Nature* 246: 15–18.

Mayntz, R. 2004. “Mechanisms in the analysis of social macro-phenomena.” *Philosophy of the Social Sciences* 34: 237–259.

Milinski, M. 1987. “TIT FOR TAT in sticklebacks and the evolution of cooperation.” *Nature* 325 (6103): 433–435.

Nagel, E. 1935. “The Logic of Reduction in the Sciences.” *Erkenntnis* 5: 46–52.

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

Nagel, E. 1970. “Issues in the Logic of Reductive Explanations.” In H. E. Kiefer and K. M. Munitz (eds.). *Mind, Science, and History*. Albany: SUNY Press, 117–137.

O'Malley, M. A. 2013. "When integration fails: prokaryote phylogeny and the tree of life." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 44 (4): 551–562.

Odenbaugh, J. and Alexandrova, A. 2011. "Buyer beware: robustness analyses in economics and biology." *Biology and Philosophy* 26: 757.

Okasha, S. 2006. *Evolution and the Levels of Selection*. Oxford: Oxford University Press.

Popper, K. 1976. *Unended Quest: An Intellectual Autobiography* (2002 ed.). London and New York: Routledge.

Popper, K. 1978. "Natural Selection and the Emergence of Mind." Dialectica 32 (3/4): 339–355.

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: St. Martin's Press.

Rosenberg, A. 1985. *The Structure of Biological Science*. Chicago: University of Chicago Press.

Rosenberg, A. 1992. *Economics—Mathematical Politics or Science of Diminishing Returns?* Chicago: Chicago University Press.

Rosenberg, A. 1994. *Instrumental Biology or the Disunity of Science*. Chicago: University of Chicago Press.

Rosenberg, A. 2005. "On the original contract: evolutionary game theory and human evolution", *Analyze und Kritik* 27: 137–157.

Rosenberg, A. 2006. *Darwinian Reduction*. Chicago: University of Chicago Press.

Rosenberg, A. 2009. "If economics is a science, what kind of science is it?" In H. Kincaid and D. Ross (eds.). *The Oxford Handbook of Philosophy of Economics*. Oxford: Oxford University Press, 55-67.

Schaffner, K. 1967. "Approaches to Reduction." *Philosophy of Science* 34: 137–147.

Schlaile, M. P., Veit, W., and Boudry, M. (Forthcoming). "Memes." In K. Dopfer et al. (eds.). *Routledge Handbook of Evolutionary Economics*. London: Routledge.

Skyrms, B. 1996. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Skyrms, B. 2010. *Signals Evolution, Learning and Information*. New York: Oxford University Press.

Sugden, R. 2001. "The evolutionary turn in game theory." *Journal of Economic Methodology* 8 (1): 113–130.

Sugden, R. 2009. "Credible Worlds, Capacities and Mechanisms." *Erkenntnis* 70: 3–27.

Sugden, R. 2011. "Explanations in search of observations." *Biology and Philosophy* 26: 717–736.

Taylor, P. and Jonker, L. 1978. "Evolutionarily Stable Strategies and Game Dynamics." *Mathematical Biosciences* 40: 145–56.

Van Der Steen, W. J. 1993. "Towards disciplinary disintegration in biology." *Biology and Philosophy* 8 (3): 259–275.

Veit, W. 2019a. "Model Pluralism." *Philosophy of the Social Sciences* 50 (2): 91-114. https://doi.org/10.1177/0048393119894897

Veit, W. 2019b. "Evolution of multicellularity: cheating done right." *Biology and Philosophy* 34 (34). https://doi.org/10.1007/s10539-019-9688-9

Veit, W. 2019c. "Modeling Morality." In L. Magnani et al. (eds.). *Model-Based Reasoning in Science and Technology*. Cham: Springer, 83-102. https://doi.org/10.1007/978-3-030-32722-4_6

Veit, W. 2021a. "Agential Thinking." *Synthese* 199, 13393-13419. https://doi.org/10.1007/s11229-021-03380-5

Veit, W. 2021b. "Model Diversity and the Embarrassment of Riches." *Journal of Economic Methodology* 28 (3): 291-303.

Von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton University Press.

Weibull, J. 1995. *Evolutionary Game Theory*. Cambridge: MIT Press.

Wynne-Edwards, V. C. 1962. *Animal Dispersion in Relation to Social Behavior*. London: Oliver and Boyd.

Croatian Journal of Philosophy
Vol. XXIII, No. 67, 2023
https://doi.org/10.52685/cjp.23.67.3
Received: April 30, 2022
Accepted: March 12, 2023

How Does Justice Relate to Economic Welfare? A Case Against Austro-Libertarian Welfare Economics

IGOR WYSOCKI and ŁUKASZ DOMINIAK*
Nicolaus Copernicus University, Toruń, Poland

This paper argues—contra some Austro-libertarians—that whether a given exchange is welfare-enhancing or welfare-diminishing does not depend on whether that exchange is just or unjust, respectively. Rather, we suggest that in light of our two thought experiments, Austro-libertarianism has at least a pro tanto reason to conceive of justice and welfare as two logically distinct ideals. This would in turn, most interestingly, predict the possibility of (a) just but welfare-diminishing exchanges and (b) unjust but welfare-enhancing ones. Upon considering possible rejoinders to our points, we suggest that Austro-libertarians abandon a justice-based notion of welfare.

Keywords: Austrian welfare economics; demonstrated preference; libertarianism; justice; voluntariness; free market.

1. *Introduction*

According to Austro-libertarians,[1] the free market is conceived in terms of property rights. Most characteristically, the main Austro-libertarian

* This research was funded in whole or in part by the National Science Centre, Poland, grant number 2020/39/B/HS5/00610. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

[1] The two welfare theorems discussed below is the crux of the Rothbardian welfare economics, which was followed—with some minor twists—by numerous Austrians: e.g. Hoppe (1990), Gordon (1993), Herbener (1997, 2008) or Hülsmann (1999). To avoid the tediousness of our prose, we shall henceforth refer to them simply as *Austro-libertarians*. However, in all fairness, we cannot but mention that

argument for the free market regime is of moral rather than economic nature. To see that consider, for example, the following quote from Rothbard ([1973] 2006: 48–49):

> It so happens that the free-market economy, and the specialization and division of labor it implies, is by far the most productive form of economy known to man, and has been responsible for industrialization and for the modern economy on which civilization has been built. This is a fortunate utilitarian result of the free market, but it is not, to the libertarian, the prime reason for his support of this system. That prime reason is moral and is rooted in the natural-rights defense of private property we have developed above. Even if a society of despotism and systematic invasion of rights could be shown to be more productive than what Adam Smith called 'the system of natural liberty', the libertarian would support this system. Fortunately, as in so many other areas, the utilitarian and the moral, natural rights and general prosperity, go hand in hand.

Here, Rothbard makes it most explicit that the "prime reason" why libertarians support the free market is moral. The fact that the regime under consideration happens to be the most productive sort of economy is only "a fortunate utilitarian result." That property rights are central to the Austro-libertarian understanding of the free market (or capitalism) and various other institutions is further evinced by the following citation from Hoppe ([1998] 2010:18):

> Next to the concept of action, *property* is the most basic category in the social sciences. As a matter of fact, all other concept […]—aggression, contract, capitalism and socialism—are definable in terms of property: *aggression* being aggression against property, *contract* being a nonaggressive relationship between property owners, *socialism* being an institutionalized policy of aggression against property, and *capitalism* being an institutionalized policy of the recognition of property and contractualism.

However, even though Austro-libertarians at large endorse the free market regime primarily because it respects property rights, they also set themselves an additional task of proving that it is the free market that always increases social utility. They want to achieve it by resorting to the concept of demonstrated preference and the Unanimity Rule. The concept of demonstrated preference refers to the actual choice that "reveals, or demonstrates, a man's preferences; that is, that his preferences are deducible from what he has chosen in action." (Rothbard [1956] 2011: 290) On the other hand, the Unanimity Rule has it that "[w]e can only say that 'social welfare' (or better, 'social utility') has *increased* due to a change, if no individual is worse off because of the change (and at least one is better off)." (Rothbard [1956] 2011:

Rothbardians are not exhaustive of Austro-libertarians. After all, one can easily point to many prominent Austrians of more or less libertarian persuasion. Suffice it to say that both Mises ([1922] 1962, [1949] 1998, 2002) and Hayek ([1960] 1978) shared a broadly construed libertarian world-view. Furthermore, it is worth noting that what singles out Rothbardians is their uncompromised adherence to absolute private property rights (see: Rothbard [1982] 2002). Still, bear it in mind that non-Rothbardian Austro-libertarians are outside the scope of the present paper.

314) Briefly stated the idea is that the concept of demonstrated preference and the Unanimity Rule can show, without passing any ethical judgements,[2] that (1) the free market always increases social utility[3] and that (2) no governmental intervention can ever increase it. In other words, the above two statements have it that just exchanges are always mutually beneficial and that unjust exchanges can never be welfare-enhancing.[4] Following Kvasnička (2008: 49), we can call claim (1) the first welfare theorem and claim (2) the second welfare theorem.[5]

[2] The reason Austrians employ the Paretian Unanimity Rule—instead of the notion of Marshall efficiency—is precisely because they disown the idea of interpersonal comparisons of utility. It is Marshall's idea of efficiency, but not the Paretian, that is committed to passing a judgement on whether social utility increased or not in situations wherein one party to an exchange benefits, whereas the other loses. For an excellent analysis of Marshall efficiency, see e.g. Friedman (1990, 2000).

[3] As one anonymous reviewer observed, it must be added that it is *the ideal free market* (i.e. the one on which there are no invasions of property rights) that allegedly ensures welfare-maximization. And indeed, in his famous essay, Rothbard ([1956] 2011: 320) writes that "[t]he free market is the name for the array of all the voluntary exchanges that take place in the world." And, as remarked by Prychitko (1997: 438), "[b]y this definition, the free market excludes invasive acts." Incidentally, we are going to elaborate on the rights-based concept of voluntariness, as adopted by libertarians, in the forthcoming parts of this essay.

[4] To avoid the tediousness of the prose we shall use "mutually beneficial" as elliptical for "mutually beneficial *ex ante*" or "mutually beneficial *in expectation*".

[5] As sharply spotted by an anonymous referee, Rothbard's two welfare theorems, as dubbed by Kvasnička, do not quite coincide with two main theorems of standard welfare economics. For, in standard welfare economics, *the first theorem* has it that under such conditions as perfect information, complete markets (characterized by every single asset having a price and no or negligible transaction cost) and with consumers and firms being price takers (i.e. with nobody having market power), all market outcomes are going to be Pareto-efficient. That is to say, what the market—under the said assumptions—is going to lead to is the situation wherein we cannot render *anybody* better off without simultaneously rendering *somebody else* worse off. This, as might be noted, bears some resemblance to Rothbard's first welfare theorem, which, recall, has it that the free market always increases social utility. However, the two theorems differ in one crucial respect. After all, Rothbard ([1956] 2011) is interested in *Pareto-superior moves* (i.e. the exchanges benefitting at least one party, while not decreasing anybody's well-being) rather than in the state of Pareto-efficiency, in which no further mutually beneficial exchanges are possible. Still, Rothbard's second welfare theorem is completely dissimilar to the second theorem of standard welfare economics. The *second welfare theorem* of mainstream economics, on the other hand, turns the first one around. To wit, whereas the first welfare theorem submits that any market allocation is Pareto-efficient, the second welfare theorem says that any Pareto-efficient allocation can be achieved by the market under the same set of assumptions as the ones under which the first welfare theorem holds. Or, more technically, for all x's, x being a Pareto-efficient outcome, there exists y, y being a distribution of initial endowments, such that the market will bring about x, given y. Needless to say, these considerations are totally unrelated to the Rothbardian second welfare theorem. Incidentally, for an excellent elaboration on the fundamental theorems of standard welfare economics, see e.g. Greenwald and Stiglitz (1986).

The present paper argues against both of these claims. Additionally, it makes a positive argument for market inefficiencies and mutually beneficial injustices, and hence for the position that justice and welfare should constitute two independent ideals within the Austro-libertarian framework. This in turn predicts that there can indeed be (a) just but welfare-diminishing exchanges and (b) unjust but welfare-enhancing ones.

The agenda of the present paper is as follows. Section 2 produces a thought experiment attempting to demonstrate that there are indeed such exchanges that should be most aptly classified as just but welfare-diminishing. Section 3, by contrast, introduces another thought experiment designed to show that we can conceive of unjust but welfare-enhancing exchanges. We believe that the said two imaginary scenarios do no violence to original Austro-libertarian methodological tools of demonstrated preference and the Unanimity Rule so that no questions are begged. Section 4 preempts possible rejoinders to our position. Section 5 concludes.

2. *Just exchanges are not necessarily welfare-increasing*

Let us start with a paradigm example of exchanges that are both unjust and welfare-diminishing. We believe that the classical highwayman's proposal "Your money or your life" is such an example. This proposal has the following biconditional structure:

Highwayman

(1) If you pay me (demand), I won't kill you (relative benefit).
(2) If you don't pay me (refusal), I will kill you (threat).

First, since the threat element promises an action that would violate the recipient's rights, the actor's payment would result in an unjust distribution.[6] Second, the payment, although unjust, is an exchange, that is, an instance of action. Finally, it is intuitively obvious that the recipient's welfare diminishes *ex ante* by paying under these conditions. Now the crucial point is that this welfare-diminishment cannot be relative to what would have happened, had he failed to pay. Since his payment was an action, he must have benefited relatively to the option foregone, that is, being killed. Otherwise, he would not have paid but rather been killed.[7] Hence, his welfare-diminishment cannot be understood in relative terms but must be explained in accordance with an

[6] That illegitimate threats result in unjust outcomes and that legitimate threats result in just outcomes is most clearly evidenced by Block's (2013) treatment of blackmail.

[7] To this effect says Mises ([1949] 1998: 351): "First, valuing that results in action always means preferring and setting aside; it never means equivalence." From this statement we can deduce that in Highwayman, as long as the highwayman's victim chooses to pay, he must prefer parting with the money to being killed.

absolute baseline. In other words, the recipient's welfare diminished compared to the situation wherein the highwayman would have nothing to do with the recipient at all. This sort of comparison involved in the idea of absolute welfare-diminishment makes perfect intuitive sense: the recipient seems to be rendered worse off when compared to the situation in which the gunman would have nothing to do with the recipient at all. It therefore follows, interestingly, that the fact that the recipient benefits relatively by handing the money to the gunman seems to be irrelevant to the estimation of his overall welfare. Since the victim's welfare obviously diminishes, welfare-diminishment cannot be explained in terms of relative benefits but must be explained in absolute terms of what would have happened if the highwayman had had nothing to do with the victim. Hence, *Highwayman* clearly represents an exchange which is both unjust and welfare-diminishing even if it, being an action as it was, benefited the victim relatively.

To provide a still more informative context, let us also consider a paradigm exchange which is just and welfare-enhancing at the same time. Suppose that the car dealer makes the following proposal to the customer:

Car Dealer

(1) If you pay me (demand), I will sell you a car (relative benefit).
(2) If you don't pay me (refusal), I will not sell you a car (threat).

First of all, since the threat element promises an action that would not violate the recipient's rights, the customer's payment would result in a just distribution. Second, since the payment is an action, the customer must have benefited relatively by paying. Otherwise, he would not have paid. However, contrary to the above example with the highwayman, the customer also benefited in absolute terms because he would have been worse off when compared to the situation in which the car dealer had had nothing to do with him at all. This exchange would therefore be just and welfare-enhancing in both senses of welfare-enhancement, that is, in relative and absolute sense.

Having spelled out crucial characteristics of, on the one hand, a paradigm case of unjust and welfare-diminishing exchanges and just and welfare-enhancing ones on the other, we are in a position to introduce our first thought experiment. Suppose that a blackmailer makes the following proposal to the blackmailee:

Blackmail

(1) If you pay me $1.000.000 (demand), I will let your reputation remain untarnished (relative benefit).
(2) If you don't pay me (refusal), I will gossip about your secrets (threat).

First of all, since the threat element promises an action that would not violate the criminal's rights, the blackmailee paying the blackmailer $1.000.000 would result in a just distribution. To see that, consider the following assessment of justice of blackmail proposals by Block (1999: 124), who has it that in blackmail scenarios "a valuable consideration is demanded, under the threat of doing something entirely licit, something that everyone would agree is legitimate if it occurred in any other context." Moreover, our author also notes that under blackmail "money is usually the valuable consideration demanded" and that "the threat is to engage in entirely legal gossip."[8]

Second, since the blackmailee paying the blackmailer is an action, the blackmailee must have benefited relatively by transferring money. Otherwise, he would not have paid. However, contrary to *Car Dealer*, the blackmailee did not benefit in absolute terms because he would have been better off when compared to the situation in which the blackmailer had had nothing to do with him at all (since then he would preserve his reputation for free). Thus, in this respect, the blackmailee is in the same position as the highwayman's victim in *Highwayman*. That is, he benefits only relatively but not absolutely. The only relevant difference between the two cases is justice of the threat element and, therefore, of the subsequent distribution. Hence, blackmail exchanges would be just, although welfare-diminishing in the relevant sense. Thus, we have a case that seems to run counter to Rothbard's first welfare theorem that just exchanges always increase social utility.

To illuminate further why we contend that blackmail exchanges do not increase blackmailees' welfare, we should come back to our distinction between benefitting relatively and benefiting absolutely. We might also call benefiting relatively benefiting in a weak sense, whereas benefiting absolutely benefiting in a strong sense. Now let us define benefitting in a weak sense as maximizing one's welfare under a newly imposed budget constraint. In fact, little wonder this sense of benefitting is weak. For we should bear in mind that every instance of human action benefits its doer at least in the weak sense. Whatever economic agents do, they maximize their expected welfare under the occurrent circumstances, whether welcome or not. However, were Austro-libertarians to adopt the weak sense of benefitting in their defence of the presumed social-welfare-enhancing character of blackmail exchanges, they would at the same time prove too much. For then, it would transpire that the gunman's proposal "Money or your life" is welfare-enhancing too. After all, whatever the gunman's victim happens to choose under the thus imposed constraint will automatically increase his expected welfare.

[8] Moreover, in this context it is worth remembering that for libertarians, the justice of arising distributions depends on the legitimacy of antecedent proposals and whether proposals are legitimate or not depends solely on the legitimacy of the threat element (see e.g. Block 2013). It is for that reason that libertarians would find the distribution of endowments arising after the blackmailee's buying off the blackmailer just.

In other words, the victim can still benefit relatively, even in such dire straits. Yet, it is a matter of course that no Austro-libertarians would be ready to bite the bullet and thus concede that the victim's exchange with the gunman constitutes a Pareto-superior move. Besides being extremely counterintuitive, this move would violate the second welfare theorem, which has it that no unjust exchanges ever increase social utility. Hence, the exchange under consideration is correctly believed to amount to a paradigm case of welfare diminishment. But if so, then, clearly, we are not warranted in inferring welfare-enhancement from the fact of benefitting relatively. In fact, as the doctrine of opportunity cost attests, benefitting relatively sweeps over the whole realm of human action and as such it is, of course, powerless to distinguish between welfare-enhancing and welfare-diminishing exchanges.

Therefore, since resorting to the weak sense of benefitting can in no way be supportive of the claim that blackmail proposals increase social utility, what is left to show is that they do not increase welfare in absolute terms. To see that, let us remind ourselves that benefiting absolutely is benefiting strongly, that is, not only given the constraint on the actor but also compared to the situation in which the constraint-maker had nothing to do with the actor. Thus, to establish whether the blackmailee actually benefits from the blackmailer's proposal we should compare this situation to a merely possible situation in which the actual blackmailee does not have to deal with the actual blackmailer at all, everything else equal. It seems quite clear that the actual blackmailee would be better off if no blackmailer were around, for in this situation the former would not even have to pay to preserve his reputation. By contrast, once the blackmailer appears on the stage and makes his blackmail proposal, there is no chance for the blackmailee to preserve his good reputation *and* keep the money. Therefore, it stands to reason that the blackmailee does not benefit absolutely when given a blackmail proposal. And, rather unsurprisingly, the same remark applies to *Highwayman*. The highwayman's actual victim would have been better off had he had nothing to do with the highwayman at all in the first place. Once confronted by the highwayman, the victim can no longer preserve his money and his life.

To summarize, since the idea of benefitting relatively may be rightly discarded as a criterion of welfare-enhancement and because the comparison involved in the notion of benefitting absolutely shows that blackmail proposals are welfare-diminishing, Rothbard's first welfare theorem seems to be challenged. For blackmail proposals, while being welfare-diminishing, are clearly just, as additionally admitted by Austro-libertarians themselves. Hence, the entire argument put forward in this section can be reduced to the following *modus tollens* reasoning. If the first welfare theorem is true, then blackmail proposals, being just as they are, are welfare-enhancing. However, blackmail proposals are not welfare-enhancing. Therefore, the first welfare theorem is false. Additionally, the blackmail proposal, although opposite morally, is economically analogous to the highwayman's proposal because both proposals make their respective recipients lose in the absolute sense. That is why, Austro-libertarians are caught in a dilemma. If they want to preserve the intuition that the highwayman's proposal is welfare-diminishing, then they are committed to regarding blackmail proposals as welfare-diminishing too, which in turn would run against the first welfare theorem since blackmail proposals are—by their lights—just. If, on the other hand, they wanted to deem blackmail proposals welfare-enhancing in order to preserve the first welfare theorem, then they would be committed to regarding the highwayman's proposal welfare-enhancing too, which would in turn run against the second welfare theorem since the highwayman's proposal is unjust.

3. *Unjust exchanges are not necessarily welfare-diminishing*

As already mentioned, the so-called Rothbard's "second welfare theorem" (see: Kvasnička 2008: 49) has it that "*no act of government whatever can ever increase social utility*" (Rothbard [1956] 2011: 323). In order to show that it is not the case, let us propose the following thought experiment, which is designed to illuminate a possibility of there being unjust and yet welfare-enhancing exchanges.

Fridge

Suppose A has an old broken fridge in his backyard, which is an economic bad for him. He would like to get rid of it, but it takes disposing of it in a faraway junkyard. Selling it would also be burdensome for him due to high transaction costs. So, the fridge just sits there in the backyard spoiling its owner's view. One day he sees, to his delight, a thief absconding with the fridge. Having realized his fridge is thus being removed for free, he decides not to interfere.

First of all, this exchange of an old fridge for the satisfaction of having it removed is unjust. Clearly, our thought experiment stipulates that person A holds a property right in the fridge. Additionally, the above scenario assumes that A has never waived his ownership rights. However, there is a worry that the putative theft cannot count as right-violating simply because A welcomes it, which might translate into a tacit waiver. But this charge is unavailable for Austro-libertarians, who repudiate the juridical significance of tacit or implicit consent.[9] As

[9] The following citation from Hoppe (2006: 389–390) is most representative: "Orthodox, i.e., statist, political theorists, from John Locke to James Buchanan and John Rawls, have tried to solve this difficulty through makeshift "tacit," "implicit," "conceptual" agreements, contracts, or state constitutions. All of these characteristically tortuous and confused attempts, however, have only added to the same unavoidable conclusion drawn by Rothbard: That is impossible to derive a justification of government from explicit contracts between private property owners, and hence, that the institution of the state must be considered unjust, i.e., the result

pointed out by Williamson M. Evers (1977: 193), the notion of tacit consent "is an overbroad extension of consent that makes it meaningless as a criterion of legitimacy." To this effect Evers quotes Gough (1957: 139), who commenting on John Locke's idea of tacit consent supporting the creation of government says that "[i]f consent could be watered down like this, it would lose all value as a guarantee of individual liberty, and the most outrageous tyrant could be said to govern with the consent of his subjects." Thus, we are justified in concluding that the exchange analyzed in our thought experiment is illegitimate, for resorting to the idea of tacit consent in order to claim that there was a tacit waiver of the fridge owner's rights is blocked for Austro-libertarians.[10]

Second of all, the above thought experiment assumes that the exchange in question involves an action on the part of A. After all, A omitted to interfere with the process of stealing and as Mises ([1944] 1998: 13) famously contented, all omissions are actions:

of moral error". See also Nozick (1974: 287) saying that "tacit consent isn't worth the paper it's not written on". Additionally, see Rothbard ([1982] 2002: 164–166); Barnett (1986: 317); Evers (1977). However, see the caveat in the footnote below.

[10] At this point, an anonymous referee made an ingenious point trying to reduce our argument ad absurdum. For, as he or she claims, if libertarians indeed do not recognize tacit consent at all, why shaking somebody's hand without his or her *explicit* consent should not count as right-violating too? In other words, would not a handshake without explicit consent be involuntary? However, clearly, libertarians would not like to deem a handshake without explicit consent involuntary? But if so, this indeed calls for making room for tacit consent at least in some situations. But then, a critical problem arises: if we do concede that libertarians must recognize tacit consent in some situations, why should our *Fridge* not involve tacit consent too? What can we offer at this point is, first, the observation that—as it follows from our examples including the alleged insignificance of tacit consent—libertarians do not recognize the legitimacy of tacit consent when it is the government that is apparently consented to. Second, we believe that it is social conventions that help us establish whether consent is given or not. For instance, libertarians would accept that a person entering a taxi and saying "Take me to the city centre" *agrees* to pay upon arrival. They would also concur that a person ordering coffee in a café agrees to pay upon drinking it. Moreover, and crucially, we contend that once we take heed of *social conventions* we should conclude that our *Fridge* scenario does not involve tacit consent, as keeping unused things in one's backyard does not *conventionally* communicate that one is ready to give up one's ownership of the said items. And it is for that reason that another person's taking of the fridge counts as a theft rather than original appropriation. Now a few words are due about the referee's counterargument involving a handshake as allegedly right-violating. We submit that whether a handshake amounts to a right-violating act again depends on the context. If two friends meet, then, most certainly, their shaking hands would be a voluntary act as the tacit consent to shake each other's hands holds between the two, as they are, after all, *friends*. However, if a man menacingly approached a woman from behind and shakes her hands, we would not be warranted in speaking of the woman tacitly consenting to such a handshake. Rather, this sort of a handshake would constitute nothing short of an act of battery. For an illuminating analysis of how social conventions are evidentiary of whether consent was given or not, see e.g. Husak and Thomas (1992).

> Praxeology consequently does not distinguish between "active" or energetic and "passive" or indolent man. The vigorous man industriously striving for the improvement of his condition acts neither more nor less than the lethargic man who sluggishly takes things as they come. For to do nothing or to be idle are also actions, they too determine the course of events. Wherever the conditions for human interference are present, man acts no matter whether he interferes or refrains from interfering. He who endures what he could change acts no less than he who interferes in order to attain another result. A man who abstains from influencing the operation of physiological and instinctive factors which he could influence also acts. Action is not only doing but no less omitting to do what possibly could be done.

Moreover, due to the fact that A acted, our thought experiment sidesteps the so-called "fallacy of psychologizing" (Rothbard ([1956] 2011: 296). For by acting in the form of omitting, he thereby demonstrates his preference for non-interference over interference. Whatever the reason A is now acting on, it remains apodictically true that, everything considered, A prefers getting his fridge stolen to intervening and thus preventing the thief from taking possession of it.

Third, since A acted, he must have benefited relatively. That is, given the thief's presence, A prefers non-interference with his fridge being stolen over being stuck with it in his backyard. But more interestingly, A also benefited in absolute terms because if there were no thief around, A would still be stuck with his fridge. Hence, we should conclude that the exchange in question was welfare-enhancing. Since it was also unjust, it follows that it then constitutes a counterexample to Rothbard's second welfare theorem.

Finally, what is important to note is that our thought experiment is also true to the Unanimity Rule, adopted by Rothbard. This is a crucial issue, for if we were to find out that at least one party to the above exchange were rendered worse off, the determination of whether the exchange was on balance welfare-enhancing or welfare-diminishing would have to rely on the interpersonal comparison of utility—an anathema to Austro-libertarians. Yet, our thought experiment seems to escape unscathed in this respect too. Clearly, the thief seems to maximize his welfare at least in expectation when he is stealing A's fridge as compared to anything else he saw as a possibility.

4. *Involuntariness charge*

Trying to put ourselves in Austro-libertarians' shoes, we can think of one truly critical objection to our position. It is for this reason that we are going to attempt to preempt it. The objection in question appeals to the notion of voluntariness, as understood by Austro-libertarians. Thus, let us first clarify what this understanding is. As Nozick famously put it (1974: 262):

> Whether a person's actions are voluntary depends on what it is that limits his alternatives. If facts of nature do so, the actions are voluntary. (I may voluntarily walk to someplace I would prefer to fly to unaided.) Other

> peoples' actions place limits on one's available opportunities. Whether this makes one's resulting actions non-voluntary depends upon whether these others had the right to act as they did.

Following Nozick, we can say that Austro-libertarians' understanding of the notion of voluntariness is rights-based. To put it simply, if A constrain~~t~~s B's options legitimately (i.e. while violating no rights of B's), B reacts voluntarily. If, by contrast, A constrain~~t~~s B's opportunity set illegitimately (viz., while violating B's rights), B reacts involuntarily. As a consequence of this theory, for example, since libertarians, as we remember, consider blackmail proposals morally permissible, they view the act of buying the blackmailer off as a voluntary payment.[11] On the other hand, since Austro-libertarians deem extortion or robbery proposals (e.g. "Give me your money or I will kill you") morally impermissible, they would deem such payments involuntary. This Austro-libertarian idea of rights-based voluntariness is further evidenced by, for example, the following quotations from Rothbard ([1956] 2011: 320) for whom, on the one hand, "[t]he free market is the name for the array of all the voluntary exchanges that take place in the world" while "rooted in the natural-rights defense of private property," on the other (Rothbard 2006: 48). After all, for Rothbard (2006: 50) the very idea of freedom is rights-based. As he points out, "Freedom is a condition in which a person's ownership rights in his own body and his legitimate material property are not invaded, are not aggressed against.... Freedom and unrestricted property rights go hand in hand."

Hence, having at their disposal this rights-based idea of voluntariness, Austro-libertarians could maintain that, for example, *Highwayman* involves the involuntary exchange due to the fact that the highwayman's victim's property rights are violated. Now since the highwayman exchange is involuntary, Austro-libertarians could try to claim that the reason for which it is welfare-diminishing is not the fact that the victim loses in absolute terms but exactly the fact that it is involuntary.[12] By

[11] See, for example, Rothbard's ([1962] 2009: 183) *Man, Economy, and State*, in which he says: "Similarly, blackmail would not be illegal in the free society. For blackmail is the receipt of money in exchange for the service of not publicizing certain information about the other person. No violence or threat of violence to person or property is involved." See also Block (2013).

[12] At this point, we would like to reassure the reader that for Rothbardians the standard of welfare-enhancement (and welfare-diminishment) is indeed justice-based rather than being rendered better off or worse off in *absolute* terms, respectively. First of all, consider the original Rothbardian ([1956] 2011: 320) attempt to argue for the free-market efficiency: "Let us now consider exchanges on the free market. Such an exchange is voluntarily undertaken by both parties. Therefore, the very fact that an exchange takes place demonstrates that both parties benefit (or more strictly, *expect* to benefit) from the exchange. The fact that both parties chose the exchange demonstrates that they benefit. The free market is the name for the array of all the voluntary exchanges that take place in the world. Since every exchange demonstrates a unanimity of benefit for both parties concerned, we must conclude that *the free market benefits all its participants*." However, as we remember, Austro-libertarians adhere to the Nozickian (1974: 262) rights-based

contrast, in the blackmail scenario, the agreement on the part of the blackmailee secured by the blackmailer's proposal is voluntary since there is no right violation looming in the case of the blackmailer spreading the unwelcome gossip. In other words, in the blackmail scenario, the blackmailer's threat is legitimate and it is for this reason that when the blackmailee agrees to pay, he does so voluntarily. Now because he agrees voluntarily, Austro-libertarians could try to argue that the exchange is welfare-enhancing, regardless of the fact that he loses in absolute terms. This sort of retort would not only establish an important difference between *Highwayman* and *Blackmail* in terms of their respective social utility but would also save the first welfare theorem against our thought experiment by showing that blackmail exchanges are both just and welfare-enhancing. The same criticism would of course apply to other cases considered in the present paper.

On the face of it, the critique pointing to the involuntary and voluntary character of the scrutinized exchanges, respectively, appears to be formidable. After all, it might seem to be the voluntariness of an exchange that secures mutual benefits, whereas the involuntariness of an exchange might be presumed to bring losses to at least one party. To appreciate it even more, we should yet again take heed of the fact

understanding of voluntariness. To reiterate, an exchange is deemed involuntary when it involves a right violation, whereas it is regarded as voluntary when it is rights-respecting. Couple those insights with all voluntary exchanges being mutually beneficial and all involuntary exchanges involving losses to (at least) one party and we end up with the ultimate standard of exchanges being mutually beneficial or not. That is, in final analysis, some exchanges are mutually beneficial by virtue of there being *just*, whereas some other exchanges are not mutually beneficial by virtue of their being *unjust*. Moreover, the idea that the welfare-enhancing and welfare-diminishing character of exchanges derives from their being just and unjust, respectively, is even more explicitly stated in Herbener (2008: 61), who has it that "[v]oluntary and involuntary interactions are defined in economics to recognize the distinction between cases in which it is possible to deduce that a person is better off from an interaction with another person and cases in which it is possible to deduce that he is worse off. Each person comes to an exchange with his naturally-owned property. A voluntary exchange occurs when neither trader uses or threatens violence against the property of the other. If the two persons trade the ownership of property without aggressive violence, then the exchange is voluntary. Given their natural ownership of property, each person chooses an alternative he prefers more than the non-interaction alternative. Both traders benefit. If one person violently aggresses against the property of the other person, then the exchange is involuntary. Given their natural ownership of property, the aggressor chooses an alternative that he prefers more than the non-interaction alternative and the victim is forced to choose an alternative that he prefers less than the non-interaction alternative. The aggressor benefits and the victim loses." Clearly, since mutual benefits depend on voluntariness of an exchange, and since the exchange is voluntary due to its rights-respecting character, then, in the end, mutual benefits are attributed to the *just* nature of the exchange. The same reasoning applies, *mutatis mutandis*, to the exchanges failing to be mutually beneficial. Ultimately, their failing to be mutually beneficial is due to their being *unjust*. Needless to say, this justice-based standard of welfare-enhancement and welfare-diminishment has nothing to do with being rendered better off or worse off in absolute terms, respectively.

that the concept of voluntariness, as employed by Austro-libertarians, is rights-based. Moreover, it must also be borne in mind that the free market, which libertarians are so keen on defending, is first and foremost about respecting rights. Hence, all these arguments combined might support the conclusion that market exchanges are mutually beneficial because they are rights-respecting and therefore voluntary, whereas all non-market exchanges are not mutually beneficial because they are rights-violative and therefore involuntary. This argument would obviously run counter to our position.

To further elucidate how the above argument could contradict our position, let us represent it in a syllogistic form:

(1) All rights-respecting (market) exchanges are voluntary exchanges
(2) All voluntary exchanges are mutually beneficial
(3) Therefore, all rights-respecting (market) exchanges are mutually beneficial

And, *mutatis mutandis*, the argument goes analogously for involuntary exchanges:

(1) All rights-violating (non-market) exchanges are involuntary exchanges
(2) All involuntary exchanges fail to be mutually beneficial
(3) Therefore, all rights-violating (non-market) exchanges fail to be mutually beneficial

However, against the above reasoning we can point out that if an exchange's rights-respecting (or market) character implies its voluntariness, and if its voluntariness in turn guarantees mutual benefits, then in the end, it is the rights-respecting character of an exchange that guarantees mutual benefits. Therefore, it seems that we do not have two separate cases for the free market but only one, that is, the case based on the rights-respecting character of the free market. After all, the fact that the free market increases welfare ultimately depends on the fact that it is rights-respecting. But remember, Austro-libertarian ambition was to make two independent cases for the free market, not one. Thus, that would mean that they failed to argue for the free market on two counts: moral and economic. After all, as pointed out by Rothbard (2006: 48–49), it is only "a fortunate utilitarian result of the free market," that it "is by far the most productive form of economy." Decidedly, it is not "the prime reason for his support of this system," for the "prime reason is moral and is rooted in the natural-rights defense of private property." Thus, if Austro-libertarians wanted to employ the above reply to our position, they would have to drop the ambition of providing two separate arguments for the free market. But that is not the only problem they would face.

For why should the standard of voluntariness precisely fit a just distribution of rights? In other words, why should only rights-respecting exchanges be voluntary and *vice versa*? Consider our fridge owner in *Fridge* yet again. Although his rights were violated due to the fridge being stolen, the fridge owner was clearly not coerced by the thief. He was not even pressurized or motivated by any sort of threat or offer. To claim that he nonetheless was coerced only due to the fact that his property rights in the fridge were violated would hardly make any sense. Moreover, he was keen on getting rid of the fridge in the first place and so he welcomed the theft. The fridge owner had a choice over interference or non-interference and given his preferences and lack of any pressure, he decided not to interfere. To resort to some formalization, our point against Austro-libertarians' idea of rights-based voluntariness as a possible counterargument to our position can therefore be expressed as the following *modus tollens* reasoning:

(1) All rights-violating exchanges are involuntary$_{\text{libertarian sense}}$ exchanges.
(2) All involuntary$_{\text{libertarian sense}}$ exchanges are involuntary exchanges.[13]
(3) The exchange in *Fridge* is a rights-violating exchange.
(4) Therefore, the exchange in *Frige* is an involuntary$_{\text{libertarian sense}}$ exchange.
(5) But, the exchange in *Fridge* is not an involuntary exchange.
(6) Therefore, it is not the case that all involuntary$_{\text{libertarian sense}}$ exchanges are involuntary exchanges.
(7) Therefore, it is not the case that all rights-violating exchanges are involuntary exchanges.

[13] We are very grateful to an anonymous referee for drawing our attention to the fact that the previous formalization of the above argument suffered from the problem of equivocation. Originally, the argument read:

All rights-violating exchanges are involuntary.
The exchange in *Fridge* is rights-violating.
Therefore, the exchange in *Fridge* is involuntary.
But, the exchange in *Fridge* is not involuntary.
Therefore, it is not the case that all rights-violating exchanges are involuntary.

Indeed, this version of our argument, as it stood, equivocated between involuntariness in the libertarian sense and involuntariness *simpliciter* or as understood pre-theoretically, or intuitively, or voluntariness as a matter of fact, if you will. It is precisely the referee's insight that enabled us to draw the distinction between involuntariness in the libertarian sense and involuntariness *simpliciter*, which in turn, we believe, rendered our argument both valid and more penetrating. Our improvement over and above the previously made argument involved adding premise (2). This particular premise states nothing short of the libertarian pretension of capturing all involuntary exchanges in terms of their rights-based standard of involuntariness, something we now call involuntariness in the libertarian sense. However, what our *Fridge* exchange is designed to show is that this particular exchange is indeed involuntary in the libertarian sense but still voluntary as a matter of fact, the observation which in and of itself is sufficient to undermine premise (2), the libertarian rights-based standard of assessing involuntariness. To wit, since we feel strongly about *Fridge* being a voluntary (as a matter of fact) exchange and since *Fridge* involves right violation, this *ipso facto* casts doubt upon the libertarian contention that all exchanges that are involuntary in the libertarian sense are involuntary *simpliciter*.

And so we cast doubt upon the premise (2) and on the libertarian pretence that all rights-violating exchanges are involuntary *simpliciter*. In other words, via modus tollens, we contend that since the argument's conclusion is implausible, the Austro-libertarian standard of rights-based voluntariness is to be jettisoned. For if the owner benefits by being deprived of his fridge, then there is probably some flaw to rights-based voluntariness. To reiterate, what rights-based voluntariness achieves is that it links (a) just exchanges with mutual benefits and (b) unjust exchanges with Pareto-inferior moves. However, both (a) and (b) were challenged by our thought experiments.

5. *Conclusion*

The aim of this paper was to argue that there is a good reason for Austro-libertarians to recognize justice and welfare as two fully distinct ideals. To bolster this claim, we launched two thought experiments designed to show the plausibility of the two types of exchange (hitherto denied by Austro-libertarians), that is, (a) unjust but welfare-enhancing and (b) just but welfare-diminishing. While proceeding with the said scenarios, we tried to do no (or as little as possible) damage to the Austro-libertarian methodological edifice. In particular, we took the original Rothbardian conceptual framework of demonstrated preference and the Unanimity Rule for granted. And it is on these grounds that we claim that the plausibility of the above-mentioned two sorts of exchanges follow. Therefore, if our thought experiments count for something, then Austro-libertarians' contention to the effect that the free market always increases social utility and that unjust exchanges never increase social utility seems unfounded.

Having thus made a *prima facie* case for unjust but welfare-enhancing exchanges and for just but welfare-diminishing ones, we tried to preempt a possible criticism appealing to the concept of voluntariness. However, as it transpired, this move is of no avail to Austro-libertarians as it simply reasserts the link between justice and welfare, something our thought experiments were supposed to undermine. And finally, given the avowed *prima facie* plausibility of our imaginary scenarios, it seems that it is the rights-based standard of voluntariness that needs revising, at least for the purposes of Austro-libertarian welfare economics.

References

Barnett, R. 1986. "A Consent Theory of Contract." *Columbia Law Review* 86: 269–321.

Block, W. 1972. "Blackmailer as Hero." *Libertarian Forum*: 1–4.

Block, W. 1999. "Blackmailing for Mutual Good." *Vermont Law Review* 1: 121–141.

Block, W. 2013. *Legalize Blackmail*. New Orleans: Straylight Publishing, LLC.

Block, W. and G. Anderson. 2001. "Blackmail, Extortion and Exchange." *New York School Law Review* 44 (3–4): 541–561.

Block, W. and R. W. McGee. 2001. "Toward a Libertarian Theory of Blackmail." *Journal of Libertarian Studies* 15 (2).

Evers, W. 1977. "Social Contract: A Critique." *Journal of Libertarian Studies* 1 (3): 185–194.

Friedman, D. 1990. *Price Theory: An Intermediate Text*. Cincinnati: Southwestern.

Friedman, D. 2000. *Law's Order: What Economics Has to do with Law and Why It Matters*. Princeton and Oxford: Princeton University Press.

Gordon, D. 1993. "Toward a Deconstruction of Utility and Welfare Economics." *The Review of Austrian Economics* 6 (2): 99–112.

Gough, J. 1957. *The Social Contract,* 2nd ed. London: Oxford University Press.

Greenwald, B. and J. Stiglitz. 1986. "Externalities in Economies with Imperfect Information and Incomplete Markets." *The Quarterly Journal of Economics* 101 (2): 229–264.

von Hayek, F. A. ([1960] 1978). *The Constitution of Liberty*. Chicago: The University of Chicago Press.

Herbener, J. 1997. "The Pareto Rule and Welfare Economics." *The Review of Austrian Economics* 10 (1): 79–106.

Herbener, J. 2008. "In Defense of Rothbardian Welfare Economics." *New Perspectives on Political Economy* 4 (1): 53–78.

Hoppe, H.-H. [1988] 2010. *A Theory of Socialism and Capitalism*. Auburn, Alabama: Ludwig von Mises Institute.

Hoppe, H.-H. 1990. "Review of *Man, Economy, and Liberty*." *Review of Austrian Economics* 4: 249–63.

Hoppe, H.-H. 2006. *The Economics and Ethics of Private Property*. Auburn: Ludwig von Mises Institute.

Husak, D. and G. Thomas. (1992). "Date Rape, Social Convention, and Reasonable Mistakes." *Law and Philosophy* 11: 95–126.

Hülsmann, J. 1999. "Economic Science and Neoclassicism." *The Quarterly Journal of Austrian Economics* 2: 3–20.

Kvasnička, M. 2008. "Rothbard's Welfare Theory: A Critique." *New Perspectives on Political Economy* 4 (1): 41–52.

von Mises, L. [1922] 1962. *Socialism. An Economic and Sociological Analysis*. New Haven: Yale University Press.

von Mises, L. [1949] 1998. *Human Action: A Treatise on Economics. The Scholar's Edition*. Auburn, Alabama: Ludwig von Mises Institute.

von Mises, L. 2002. *Liberalism in the Classical Tradition*. San Francisco–New York: Cobden Press, The Foundation for Economic Education.

Nozick, R. 1974. *Anarchy, State, and Utopia*. Oxford: Basil Blackwell.

Prychitko, D. 1997. "Expanding the Anarchist Range: A Critical Reappraisal of Rothbard's Contribution to the Contemporary Theory of Anarchism." *Review of Political Economy* 9 (4): 433–455.

Rothbard. M. [1956] 2011. "Toward a Reconstruction of Utility and Welfare Economics." In: *Economic Controversies*. Auburn: Ludwig von Mises Institute, 289–333.

Rothbard, M. [1970] 2009. *Power and Market*. In: Murray N. Rothbard, *Man, Economy, and State with Power and Market*. Auburn, Alabama: Ludwig von Mises Institute: 1047–1369.

Rothbard, M. [1973] 2006. *For a New Liberty. The Libertarian Manifesto.* Auburn: Ludwig von Mises Institute, Scholar's Edition.

Rothbard, M. [1982] 2002. *Ethics of Liberty*. New York and London: New York University Press.

Wertheimer, A. 1989. *Coercion*. Princeton: Princeton University Press.

Croatian Journal of Philosophy
Vol. XXIII, No. 67, 2023
https://doi.org/10.52685/cjp.23.67.4
Received: July 22, 2022
Accepted: April 10, 2023

Imagination, Thought Experiments, and Personal Identity

MICHAEL OMOGE
University of Alberta – Augustana, Camrose, Canada

Should we descry the nature of the self from thought experiments? Shaun Nichols says 'maybe,' but only if we use thought experiments that do not recruit the indexical "I" (non-I-recruiting). His reason is that the psychology of "I" perforce mandates that imagination responds to thought experiments that recruit it (I-recruiting) peculiarly. Here, I consider whether he is correct about non-I-recruiting personal identity thought experiments. I argue positively using the same framework, i.e., considering the underlying psychology.

Keywords: Propositional imagination; cognitive architecture; personal identity; thought experiments.

1. *Introduction*

In no area of philosophy are thought experiments more used than in personal identity, and yet, in no area are they disparaged than in personal identity. One general reason personal identity thought experiments (PITEs) are said to fail is that propositional imagination[1] (hereafter, simply as 'imagination') breaks down in them. But if so, then this breakdown of imagination in PITEs must be traceable to some

[1] Propositional imagination is a propositional attitude that has linguistically expressed content. It is often contrasted with experiential imagination, which involves consciously entertaining mental imagery. By only talking about propositional imagination here, I do not mean that experiential imagination is not involved in thought experiments, but rather that talk of cognitive architecture—which turns on drawing a similarity between imagination and belief—is often taken to mean that experiential imagination is excluded. But see Omoge (Forthcoming) for how to include it.

faults in the 'cognitive architecture' of imagination. Where cognitive architecture "is a theory about the mind at the functional—as opposed to, say, neurological or biological—level that aims to explain relevant psychological phenomena [by] (literally) drawing out the functional connections between various components of the mind" (Miyazono and Liao 2016: 234).

Shaun Nichols (2008) notices this link between the failure of imagination in PITEs and the cognitive architecture of imagination, deploying the link to expose a shortcoming in how imagination responds to PITEs that recruit the indexical "I" (I-recruiting PITEs). Nichols focuses on Bernard Williams' (1970, 1973) modification of the Lockean body-swap PITE where the psychological properties of person A are transferred to person B. Nichols' diagnosis of why imagination breaks down in this (and other) I-recruiting PITE is that at the psychological level, "I" is semantically impoverished in that it does not come with all the historical details that characterize the speaker of the I-token. He adds that while this poverty renders "I" flexible such that there are no obstacles to imagining scenarios that recruit it, the flexibility makes it possible for an agent to imagine that *I am someone else* even when their defining psychological characteristics are destroyed, which is problematic. Thus, he concludes that we should not use I-recruiting PITEs to draw metaphysical conclusions about the self. He, however, suggests that non-I-recruiting ones may be so used.

My goal in this paper is to consider whether Nichols is right about non-I-recruiting PITEs: do they succeed in leading us to what is essential about the self? It is important to consider this question because if imagination also fails in non-I-recruiting PITEs, such that they, like their I-recruiting counterparts, fail to lead us to what is essential about the self, then that would be the final nail in the coffin for PITEs in general. Philosophers would have been doing something terribly wrong by relying on them. Although things are not so straightforward, I will argue here that Nichols' optimism is warranted. Non-I-recruiting PITEs succeed in leading us to appropriate metaphysical conclusions about the nature of the self.

I begin by discussing the cognitive account of imagination Nichols relies on (Section 2). I then rehearse how he uses the account to show why we should not infer the nature of the self from I-recruiting PITEs, but that we may from non-I-recruiting ones (Section 3). In Section 4, I explain why the cognitive account of imagination Nichols relies on is not straightforwardly compatible with non-I-recruiting PITEs; so, I give an updated version. In Section 5, I show that the updated account is compatible with non-I-recruiting PITEs. In Section 6, I use this compatibility to show why Nichols' optimism about non-I-recruiting PITEs is not misplaced.

But before I begin, let me give some examples of non-I-recruiting PITEs to clarify the scope of the discussion in this paper. Non-I-recruiting PITEs include but are not limited to the original Lockean body-swap, Parfit's (1984) fission (where the brain of one of an identical triplet is split into two and put in the bodies of the two members of the triplet), Parfit's Russian (where a Russian lost his memory but he had already told his wife to share his belongings if that ever happens), Parfit's teleporter (where someone is broken down into molecules and reassembled somewhere else), and their many variants by other theorists. Though I will only focus on fission in this paper, what I will say about it will generalize to all non-I-recruiting PITEs.

2. *The cognitive architecture of imagination*

Nichols (2008) relies on Nichols and Stich's (2003) cognitive account of imagination to show why imagination behaves peculiarly in I-recruiting PITEs. According to Nichols and Stich, the cognitive architecture of imagination comprises an 'imagination box,' which is a workspace and storage unit where imaginings are temporarily stored and manipulated, a 'script elaborator' that generates and embellishes imaginings, and an 'UpDater' that enables reasoning with imaginings. For Nichols and Stich, these cognitive structures help to explain what happens when we practically imagine, for instance, in pretense and mindreading.

In pretending to have a tea party, the representation *We are going to have a tea party* is generated as the imagination premise by the script elaborator and placed in the imagination box. The content of the belief box is then (copied and) put inside the imagination box as further premises. The UpDater then filters out the beliefs that are incompatible with the imagination premise. Since what is left after this filtering would be insufficient to yield the target imagining, Nichols and Stich say that some of the unfiltered-out beliefs contain 'scripts' (e.g., a script for how tea parties typically unfold), where scripts are psychological paradigms that describe appropriate sequences of events in a particular context (Schank and Abelson 1977). Since scripts are unrestrictive—actors often go off-script, improvising their acts—the script elaborator teases out elaborations on the sequences of events detailed by scripts. For Nichols and Stich, this is how imagination operates psychologically.

One component of this account is that imagination interacts with the same inference mechanisms with which belief interacts. This, according to Nichols and Stich, is why the UpDater, which is part of our inference mechanisms, is also at work in belief episodes. We update our beliefs all the time without needing to upend everything we know. In short, imagination and belief are in *the same code*, i.e., they have the same contents and logical form, and they interact with the same inference mechanisms. Put differently, inference mechanisms will treat imagination and belief in much the same ways. Nichols (2004) calls this component of the cognitive account of imagination the 'single code hypothesis.' Nichols (2008) thinks this hypothesis holds the secret to why I-recruiting PITEs should not be used to infer the nature of the self.

3. *Nichols on I-recruiting and non-I-recruiting PITEs*

Nichols' goal is to explain why imagination responds in the way Williams (1970, 1973) describes. According to Williams, imagining the Lockean body-swap PITE from a 1st person rather than a 3rd person perspective (i.e., turning it into an I-recruiting PITE) problematizes the psychological accounts of personal identity (e.g., Parfit 1984). When imagined from the 1st person's perspective and adding the constraint that one of the swapped bodies would be tortured after the swap, Williams argues that the imaginer would lack one vital respect. The respect of knowing "what was going to happen—torture, which one can indeed expect to happen to oneself, and to be preceded by certain mental derangements as well" (1970: 168). Lacking this respect, William concludes, suggests that the imaginer survives the destruction of their psychological properties, contradicting the psychological accounts of personal identity.

Nichols thinks the reason imagination responds to I-recruiting PITEs in this way "turn on peculiar features of imagining with indexicals" (2008: 521). What peculiar features? According to him, to accommodate indexicals in psychology, an internal mental symbol that corresponds to their semantics must be postulated. Now, the semantics of indexicals is not determined by contents. People with Alzheimer's disease, for example, use "I" frequently and appropriately, even in the late stages of the disease. Likewise, you can wake up in the dark with (a temporary) total amnesia and still be able to think *I have a headache*. Rather, the semantics of indexicals "is determined [...] by the sparse character ('the speaker of this token of "I"') plus the context" (Nichols 2008: 523). Nichols calls the internal mental symbol that corresponds to this impoverished semantics of indexicals the 'I-concept.' He then argues that the I-concept is why imagination responds peculiarly to I-recruiting PITEs.

Since inference mechanisms respond to the format, not (simply the) denotation of representations (Fodor, 1987), Nichols says that inference mechanisms will respond to the format of indexical representations, i.e., the I-concept. If so, then the poverty of the I-concept explains why there is no limitation to imagining with the *I*, not even when your psychological properties are destroyed: "In particular, the fact that all of my distinctive psychological properties are gone is no obstacle whatsoever. Given the poverty of [the I-concept], there is no constraint against the representation *I exist in this location with completely different psychological properties*" (Nichols 2008: 527). But once it is clear why we can imagine with the *I*, even with different psychological properties, it becomes clearer that we must be careful about what we make of the imagined I-scenarios.

Given the single code hypothesis (Section 2), inference mechanisms interact with the I-concept in the belief context in much the same way they interact with it in the imagination context. However, in the belief context, there is no problem arising from the poverty of the I-concept. When I wake up in the dark with total amnesia, there is still a plausible sense in which I am the referent of the I-concept, perhaps because my psychological properties still subsist, although I have no conscious access to them at the time. But there is no such sense in the imagination context under discussion (i.e., I-recruiting PITEs) precisely because my psychological properties are now destroyed, and so imagining that I persist in their absence is problematic. Consequently, Nichols warns:

> Thus, it is dangerous to draw any metaphysical conclusions from these imaginative exercises with the *I*. More generally, we should be exceedingly wary of trying to descry the nature of the self through thought experiments that invoke the *I*. Imagining with the *I* sends us on wild thought experiment rides, but the resulting intuitions are likely not a reliable guide to what the self *really* is. (2008: 529, original italics)[2]

While I-recruiting PITEs may be unreliable guides to metaphysical conclusions about the self, Nichols signals that non-I-recruiting ones may fare better: "If we are to use thought experiments to assess what is and isn't essential to the self, we would do well to exclude the cases that trade on the I-concept" (2008: 529). This optimism, however, will not get off the ground unless some of Nichols' other commitments are addressed.

4. *Metaphysical modality and the cognitive architecture of imagination*

Elsewhere (Nichols 2006a), Nichols argues that Nichols and Stich's cognitive account shows that imagination is an unreliable guide to metaphysical modality. Given the single code hypothesis, which suggests that inference mechanisms will balk at contradictions in the belief context, it follows that they will also balk at contradictions in the imagination context. This, Nichols says, is why we face imaginative blocks when we attempt to imagine metaphysical impossibilities,[3] leading him to the conclusion that imagination is an unreliable guide to metaphysical modality. Imagination's natural domain is practical modalizing (e.g., pretense), not metaphysical modalizing (e.g., personal identity).

But if so, then non-I-recruiting PITEs, like their I-recruiting counterparts, will become unreliable guides to metaphysical conclusions about the self as well, although for different reasons. Where I-recruiting PITEs are unreliable because the psychology of the *I* does not mix

[2] Outside PITE, Williams also raises a puzzle for imagining in the 1st person perspective—namely, why is it much easier to imagine that *I am Napoleon* than imagine that *Someone else is Napoleon*? Nichols also responds to this puzzle. I will say something about his response later in Section 6.

[3] Beyond metaphysical modalizing, Nichols also uses the same argument to explain why we face imaginative resistance in fiction (Nichols 2004, 2006b), and why we face difficulty in imagining our own nonexistence (Nichols 2007).

well with imagination, non-I-recruiting ones will be unreliable because they are not the natural domain of imagination. In short, as things stand, non-I-recruiting PITEs are not compatible with the cognitive architecture of imagination. Thankfully, Omoge (2021) has shown that Nichols' skepticism about using imagination to metaphysically modalize is unwarranted. Though his argument is layered, I will recap the relevant aspects here, and together with a caveat I will add later, I will argue that non-I-recruiting PITEs are not incapacitated by the cognitive architecture of imagination.

Central to Omoge's view is ascribing a larger role to scripts than Nichols and Stich do. He argues that scripts are (1) activated conceptually given the imaginer's theoretical assumptions such that a script type is rarely similarly tokened by two imaginers and (2) often compositional given the debate's etiology such that the manner of their composition explains how the imaginers get different imaginative outcomes. For instance, when Chalmers (1996) says zombies are possible, and Shoemaker (1999) says they are impossible, not only do they each token different zombie scripts, their differently tokened script explains their different individual stances. Since Chalmers says human actions are decomposable into phenomenal and functional descriptions, his zombie script decomposes into scripts for those descriptions such that his phenomenal action script leads him to the possibility of zombies. Since Shoemakers says human actions are both phenomenal and functional, his zombie script does not so decompose, and so it can only lead to the impossibility of zombies.

Omoge also foregrounds Schank and Abelson's notion of 'interference' to account for the correct usage of imagination in metaphysical modalizing. Where interferences are mental states that prevent the normal unfolding of a script and which often sneak into the imaginative process during the composition of scripts. For instance, in reasoning his way to how functional properties fail to neatly supervene on phenomenal ones, Chalmers may have made some invalid reasoning steps, such that there are some interferences lurking in his zombie script. If so, then he would have wrongly used imagination to reach his view that zombies are possible.

Omoge thinks that due to theoretical assumptions, interferences often go unnoticed, and so are left uncorrected, and even when pointed out, the involved theories may make the imaginer resolute. This, he says, shows that the psychology of imagination and metaphysical modality come apart. For we now have an account of how an agent's imaginative processes can be faulty, which says nothing about the metaphysical conclusions they arrive at via imagination—after all, Chalmers could also use another cognitive faculty, e.g., intuition, to reach the same conclusions, and, certainly, the cognitive architecture of imagination is not identical with that of intuition.

Lastly, Omoge gives an evolutionary psychological argument against Nichols' skepticism about the usage of imagination in metaphysical modalizing. In his view, talk of a natural domain matters little, if at all, because evolution does not ready-make all our cognitive faculties; some are appropriations of others. For example, spatial reasoning, which we have gone to appropriate for geometry. Omoge says the same appropriation holds for practical and metaphysical modalizing. Metaphysical modalizing may not be the natural domain of imagination, but that does not mean we are thereby barred from so using imagination. After all, geometrical reasoning is not the natural domain of spatial reasoning, yet it is indispensable. Talk of a natural domain matters little when considering the usefulness of a cognitive faculty.

While this view is commendable, Omoge does not address why Nichols is skeptical about the usage of imagination for metaphysical modalizing, which, recall, is that the single code hypothesis predicts that inference mechanisms would balk at contradictory imaginings because they balk at contradictory beliefs. I will conclude this section by supplying a rebuttal to this claim.

Here is the fact: imagination can be used to reason about contradictions, so it is factually incorrect that inference mechanisms balk when we so use imagination. In fact, it is factually incorrect that they balk at mathematical impossibilities like 1+1=7, which are the examples Nichols uses—Graham Priest (2016), for example, says he can perfectly imagine them. But he should not be able to do so if Nichols is correct. How, then, should we explain the imaginative processes of outliers like Priest? And Nichols should want to explain their imaginative processes since he says his view maps onto the cognitive architecture of imagination, which is identical for everyone. My own view is that Nichols gives up too quickly. The way out, as I see it, lies with the UpDater.

Nichols and Stich (2003: 32) set up the UpDater as though it works only in the involuntary mode, i.e., automatically. But I think it can also work in the (semi)voluntary mode. In the contexts of belief and practical modalizing, nomological laws are fundamental to how the UpDater filters out incompatible beliefs. Thus, the UpDater can work independently of what the agent wants to achieve—it just needs to follow the dictates of nomological laws, which, supposedly, are mentally filed in some determinate ways. Believing and practical modalizing are typically automated processes (Connors and Halligan 2015). You may withhold believing that your child, who was asleep in the bedroom, is the person giggling in the living room, at least until you peep to confirm, but believing so was triggered by the giggles you heard (assuming that both of you are alone in the house). Not so for metaphysical modalizing since everyone agrees that nomological laws are suspended therein. Without the guidance of the mental file for nomological laws, the UpDater falls back to what the agent wants to achieve. Simply, in metaphysical modalizing, the agent seizes control of the UpDater, telling it which beliefs to filter out, thereby making the UpDater sensitive to the agent's goal. Thus, outliers like Priest are voluntarily filtering out beliefs that would block them from imagining metaphysical impossibilities. Not everyone can do this, however; relevant beliefs are needed. I will return to this in Section 6.

This view that the UpDater is sensitive to the agent's goal is not an affront to the single code hypothesis, it must be said. Nichols and Stich only say that inference mechanisms will treat beliefs and imaginings in *much* the same way, i.e., the hypothesis admits some differences between beliefs and imaginings. Nichols (2006a) himself discusses some of these differences at length. What I am adding, then, is that the UpDater's sensitivity is another difference in how inference mechanisms treat beliefs and imaginings. In belief and practical modalizing contexts, the UpDater is not sensitive to the agent's goal, but it is in metaphysical modalizing contexts. If so, then Nichols' skepticism is indeed unwarranted because the single code hypothesis does not, in fact, show that imagination cannot lead to metaphysical modality. We only need to build sensitivity to the agent's goal into the UpDater, and the single code hypothesis will accommodate metaphysical modalizing.

Now that we have seen how the cognitive architecture of imagination can be updated to become compatible with metaphysical modalizing, we can proceed to check whether non-I-recruiting PITEs, since they are cases of metaphysical modalizing, do indeed fare better than their I-recruiting counterparts vis-a-vis the nature of the self, as Nichols suspects. First, let us demonstrate the compatibility of non-I-recruiting PITEs so as not to beg the question.

5. *Non-I-recruiting PITEs and the cognitive architecture of imagination*

As I said (Section 1), I will focus on Parfit's fission in the remainder of this paper for simplicity's sake, although what I will say is generalizable to other non-I-recruiting PITEs. In fission, identical triplets were involved in an accident such that the body but not the brain of one is damaged (Brainy), and the brains but not the bodies of the other two are damaged (Lefty and Right). Parfit asks that if Brainy's brain is split into two halves such that Lefty gets the left half and Righty gets the right half, which of Lefty and Righty will be identical to Brainy? His famous answer: neither. From this, he concludes that what matters when identity does not obtain is psychological continuity, not personal identity.

First, let me show how his conclusion is subserved by the (updated) cognitive architecture of imagination, segueing from there to whether he uses imagination correctly to arrive at the conclusion. This second task is important because if fission is to succeed in leading us to what is essential about the self, then a good starting place is whether the conclusions it affords were correctly arrived at in the first place. As we all know, an invalid conclusion cannot be sound.

In fission, the invitation to imagine that "identical triplets were involved in an accident ...” will signal to the script elaborator to generate an imagination premise, which will be put inside Parfit’s imagination box. The contents of his belief box will then be put inside the imagination box as further premises to yield the target imagining—namely, when identity does not obtain, what matters? His UpDater will then filter out any beliefs he may have that will be incompatible with the imagination premise, for example, some nomological beliefs about the physical impossibility of splitting brains into two. Here, as I said (Section 4), the UpDater is operating in a voluntary mode in that Parfit is manually controlling it, telling it to filter out the incompatible nomological beliefs, even though his UpDater will not filter the beliefs out were he not metaphysically modalizing. He can do this because he is a seasoned personal identity thinker, such that he has the relevant theoretical assumptions to maintain a coherent thought process despite manually hijacking the UpDater. For comparison, a first-year philosophy student may not be able to suspend the influence of nomological laws if they suppose that fission is possible.

Being a seasoned personal identity thinker would also enable some of Parfit’s UpDater-unfiltered-out beliefs to contain a script that details how PITEs typically proceed, i.e., he has PITE scripts or, in our case, a fission script. Like any script, this fission script will be unrestrictive in that further details about thought experiments can be teased out from it independently of Parfit’s theoretical assumptions. Simply, Parfit’s script elaborator will embellish the imaginative scenario in ways not informed by his theoretical assumptions without straying from the scope set by the fission script. Thus, from what he imports into the imaginative process—which, of course, are the UpDater-unfiltered-out beliefs—imagination will continue in an autonomous mode, fleshing out other relevant details.

Now, as we have seen (Section 4), the fission script will be activated conceptually, i.e., when key concepts like ‘personal identity’ and ‘psychological properties’ are instantiated in Parfit’s imaginative process. Since theoretical assumptions are rarely ever identical for two agents, the fission script is rarely ever identically tokened by two philosophers. Thus, when Gendler (2002) argues that Parfit is mistaken in saying that psychological continuity, not personal identity, is what matters, Gendler’s fission script differs from Parfit’s.

In addition to being activated conceptually, we have also seen that scripts are also compositional, given the etiology of the debate (Section 4). If so, then the fission script is compositional along the ‘prudential concern’ etiology of the debate. Where prudential concern, as it is used in the personal identity literature, is the sort of concern we bear towards our future selves, and prudential concern can be understood in both psychological and numerical terms. The fission script, then, is composed of a script for psychological continuity and another script for numerical identity. Since the compositionality of scripts informs different metaphysical modalizing conclusions, it follows that the manner in which the fission script is composed for Parfit and Gendler explains why they arrive at polar opposite conclusions.

Simply, given Parfit's and Gendler's theoretical assumptions and the compositionality of their fission scripts, the scripts can each unfold in ways that prioritize one of the component scripts. Parfit's theoretical assumptions guide his fission script to prioritize the script for psychological continuity. Hence he says: "In all ordinary cases, personal identity and [psychological continuity] coincide. When they diverge, [psychological continuity] is what matters. That strongly suggests that, in all cases, [psychological continuity] is what matters" (Unpublished paper, but the quote is from Gendler 2002: 44). On the other hand, Gendler's theoretical assumptions guide her fission script to prioritize the script for numerical identity. Hence she says: "The fact that two features coincide in all actual cases may mean that there is no straightforward way for us to determine how we would or should respond to either in isolation" (Gendler 2002: 35).

Now, Gendler does not just say Parfit is wrong; she also says he could not have arrived at his conclusion imaginatively. This seems to be a step too far if what I have said here is correct. As we have just seen, it is consistent with the cognitive architecture of imagination that imagination can lead different agents to different imaginative conclusions, at least insofar as each conclusion follows from the normal unfolding of the agents scripts. Both Parfit's and Gendler's polar opposite conclusions follow from the normal unfolding of their different fission scripts. Everything, so far, is by the book.

We can go a step further, however, by checking whether any of them wrongly used imagination to arrive at their respective conclusions. To do this, we only need to identify in whose fission script interferences lurk. For instance, if Gendler's argument is correct, then some interferences lurk in Parfit's fission script. According to her, Parfit wrongly thinks that because psychological continuity and numerical identity ordinarily coincide, imaginary cases where they diverge show that the former is what matters. Such an illicit move would constitute an interference, blocking the normal unfolding of Parfit's fission script, such that he would have wrongly used imagination to arrive at his conclusion. *Mutatis mutandis* for Gendler if we can isolate the interferences lurking in her fission script. We should not, however, expect that neither Parfit nor Gendler will change their view if the lurking interferences are pointed out. As I have said (Section 4), when interferences are hooked up to theories, theoretical assumptions might, and they often do, make philosophers resolute, even when lurking interferences are pointed out. Thus, if Gendler indeed points out the interferences lurking in Parfit's fission script, we should not expect that he thereby changes his mind.[4]

[4] Gendler thanked Parfit for providing comments on earlier versions of her paper in the acknowledgment section. So, there is no doubt that he read the paper, yet her arguments did not sway him. He still published numerous works between

Interferences can also be psychological, not always conceptual, as in the above, but I do not think psychological interferences pose any threat to non-I-recruiting PITEs or any imaginative exercise for that matter. It has been argued that since the laboratory of thought experiments is the mind, PITEs (as well as other kinds of thought experiments) are subject to a host of psychological biases, like seeing ourselves in positive lights (e.g., Brown 1986, Taylor and Brown 1988). Consequently, Unger (1990) says these psychological biases jeopardize the reliability of PITEs. Put in our terms, the biases would make PITE scripts unfold in different ways than they ordinarily would, and so they are interferences. They are psychological interferences.

However, unlike their conceptual counterparts, psychological interferences would easily be correctable once pointed out, suggesting that their easy correction is a function of not being hooked up to background theories. If so, then I sincerely doubt that any philosopher would refuse to account for psychological interferences in their imaginative processes once pointed out. In fact, psychological interferences are one way we improve our imaginative processes. I do not take imaginative conclusions at face value anymore; I look out for where I might have overestimated my own abilities. I am confident that this applies to Parfit and Gendler as well. In short, psychological interferences are no threat to the success of non-I-recruiting PITEs.

It might be said, following Wilson and colleagues (1994, 2002), that even if we are aware of psychological interferences, we lack access to the ongoing psychological processes, and so we cannot decontaminate in real-time. It is unclear to me, however, why such access is required, not least because, typically, psychological processes are subpersonal. Take the UpDater. In some ways, its job is to decontaminate, and typically (i.e., in the contexts of belief and practical modalizing, when it works in the involuntary mode), it does this without our awareness. When you hear someone giggling in the living room, and the UpDater updates your belief system—from "my child is sleeping" to "my child is awake"—it brackets out some psychological biases as it does so, e.g., that you are not hallucinating the giggles. If so, then talk of immediate access holds little, if any, weight in talk of decontamination. Decontamination is psychological, not phenomenological.

So far, I have argued that neither Parfit's nor Gendler's conclusion about the self is wrong, although we might be able to say which of them wrongly used imagination to arrive at their conclusion. I want to end this section by saying that there is a deeper sense in which interferences can prove fatal for a philosopher's conclusion about the self. One reason fission is popular is that it aims to show that the non-reductionist, who is committed to identity being always what matters, faces a kind of *reductio ad absurdum*. If identity is always what matters,

2002 and 2017—when he died—that propagate the same idea that what matters is psychological continuity, not personal identity.

then the non-reductionist must describe the outcome of fission in identity terms, yet any such description conflicts with some principle to which they are also committed. Brainy cannot be both Lefty and Righty given the necessity of identity; he cannot be neither, as he survives in the single case, which is no different from each side of the double case; he cannot be Lefty rather than Righty as that would make identity arbitrary.[5] Simply, whatever the non-reductionist say is wrong on their own terms. One might say then that Parfit's argument is meant to show that there is something internally wrong with the non-reductionist's fission script.[6]

I should stress that this deeper sense in which interferences are useful has not taken us too far afield. We are still within the scope of talking about the correct usage of imagination to arrive at metaphysical conclusions about the self; we have not been transported to talking about whether the conclusions themselves are correct. The latter is a metaphysical discussion; the former is a cognitive psychological one. The fact that interferences can be fatal to the success of an imaginative act is part of what "using imagination wrongly" means. Put simply, interferences do not merely reveal the thought experimenter's theoretical commitments; sometimes, they do much more, revealing why some thought experiments work and why some others do not work. If fission succeeds in leading us to what is essential about the self, then its success is at the expense of the non-reductionist. This analysis is compatible with the cognitive architecture of imagination.

What we have come to then is an explanation of non-I-recruiting PITEs with the cognitive architecture of imagination. Put plainly, it is an explanation of why imagination does not fail in non-I-recruiting PITEs. That said, such an explanation does not tell us whether the metaphysical conclusions non-I-recruiting PITEs deliver reveal anything essential about the self. After all, as Nichols points out (Section 3), we can explain I-recruiting PITEs with the cognitive architecture of imagination, but once we do so, we see why we should not infer the nature of the self from them. Thus, we must still ask whether this compatibility between the cognitive architecture of imagination and non-I-recruiting PITEs reveals the same thing about the nature of the self. Does it reveal that we should not infer the nature of the self from non-I-recruiting PITEs? I will argue that it does not.

[5] Gendler is not a non-reductionist in the sense I am using the term here. Her misgiving with Parfit is just that his explanation for why prudential concern subsists in the absence of identity is wrong: "Nevertheless, as I have maintained throughout, Parfit is right that if Brainy were to undergo fission, the relation of prudential concern he would find himself bearing to Lefty and to Righty would be rational—even if he knew that he was to undergo fission. What Parfit is wrong about is the explanation of this" (2002: 51).

[6] Thanks to an anonymous reviewer for this journal for this stronger sense in which interferences are useful.

6. *Should we descry the nature of the self from non-I-recruiting PITEs?*

Central to answering this question is the challenge that non-I-recruiting PITEs are impoverished in that they lack relevant background information, and so we should be cautious when drawing metaphysical conclusions about the self from them (Wilkes 1988; van Inwagen 1997; Schechtman 2014). For instance, Wilkes says:

> How often [do fission occur]? Is it predictable? Or sometimes predictable and sometimes not, like dying? Can it be prevented? Just as obviously, the background society, against which we set the phenomenon is now mysterious. Does it have such institutions as marriage? How could that work? Or universities? It would be difficult, to say the least, if universities double in size every few days, or weeks, or years. Are pregnant women debarred from splitting? The *entire* background here is incomprehensible (1988: 11, original italics).

The point here is that nouns (common and proper) come with all the descriptive (Russell 1911) and/or causal-historical (Kripke 1980) residuals that characterize them, which non-I-recruiting PITEs leave out. We learn names of places or things at elementary schools, names of people at their christening, or when we come to know/meet them, and we keep updating the descriptions associated with the names throughout life. Not supplying these associated descriptions, therefore, makes non-I-recruiting PITEs incomplete. Being so incomplete, we should take them with the proverbial pinch of salt, in almost the same manner we take their I-recruiting PITEs that are equally impoverished.

I agree that non-I-recruiting PITEs are descriptively impoverished in the above way, but I deny that this poverty of description amounts to anything significant. It does not amount to non-I-recruiting PITEs failing to lead us to metaphysical conclusions about the nature of the self. My reason is that this challenge (hereafter, as the Wilkes-Van Inwagen-Schectman challenge), as evident from the last sentence of the previous paragraph, wants to parallel non-I-recruiting PITEs with I-recruiting PITEs, which cannot work. The Wilkes-Van Inwagen-Schectman challenge wants to say that since I-recruiting PITEs are descriptively impoverished, which is why they are unreliable guides to the nature of the self (Section 3), so too will the descriptive poverty of non-I-recruiting PITEs make them unreliable guides to the nature of the self. This argument does not work.

The reason I-recruiting PITEs are descriptively impoverished is that the mental symbol underwriting their operation (i.e., the I-concept) is also descriptively impoverished (Section 3). This is not the case for non-I-recruiting ones. Though they are descriptively impoverished, their descriptive poverty is not caused by the descriptive poverty of the mental symbol underwriting their operation. Nichols puts this difference in psychological structure between I-recruiting and non-I-recruiting PITEs this way:

> But even if both indexicals and proper names have similarly Kripkean semantics, it would be a mistake to conclude that this means that indexical concepts and proper name concepts are also equivalent in their psychological characteristics. Rather, it's plausible that the processing associated with the I-concept differs in important ways from the processing associated with proper name concepts. To take one example, we often deploy proper names that seem nonunique, as when I think *Michael is meeting me for lunch.* I know which *Michael* I have in mind, and it's plausible that this is because of the information I have associated with that token of *Michael*. By contrast, since there's only one I-concept, I never need to worry about disambiguating it. (2008: 523)

If so, then even though both I-recruiting and non-I-recruiting PITEs are descriptively impoverished, their psychological structures differ. Call the mental symbol underwriting the operation of nouns the 'noun-concept'. Unlike the I-concept, which is flexible (Section 3), the noun-concept is rigid because it contains different mental files for different nouns. For instance, there are separate files for the many *Michaels* I know, and each file keeps getting updated as more historical facts about each of them come to my awareness. If one of them wins a Nobel, that fact will not be stored in the file of a *Michael* who is a soccer player. In short, where the I-concept is poor, the noun-concept is abundantly rich.

Here, then, is the psychological difference between I-recruiting and non-I-recruiting PITEs. Since the I-concept is poor, it is functioning normally in I-recruiting PITEs, which are also descriptively poor. This is why it is easy to imagine that *I am someone else*: the I-concept has no descriptive content, so it works anyway. I-recruiting PITEs inherit the descriptive poverty of the I-concept. Contrariwise, since the noun-concept is abundantly rich, it is not functioning normally in non-I-recruiting PITEs, which are descriptively poor. This is why it is difficult to imagine that *Obama is Napoleon*. My noun-concept has separate files for *Obama* and *Napoleon*, which contain all the historical facts I associate with them, and so the noun-concept finds it difficult to combine or crisscross data from both files. Non-I-recruiting PITEs do not inherit the descriptive wealth of the noun-concept.

The Wilkes-Van Inwagen-Schectman challenger may respond that all that this talk of malfunctioning of the noun-concept shows is that imagination also fails in non-I-recruiting PITEs, just as it fails in I-recruiting ones, such that their parallelism stands. Put differently, they would say that we are being asked to imagine a world where the mental files we have for nouns are different from the ones we currently have, but we are not told what data they contain, and this is troubling because we are using our current concepts for the nouns in the imagined world. Thus, when Parfit talks about fission, the Wilkes-Van Inwagen-Schectman challenger would retort that he skips relevant details about *brains*, *triplets*, *splitting*, and so on. As we saw, the complaint is that details like "How often do fission occur? Is it predictable? Can it be prevented?" (Wilkes 1988: 11) are skipped.

To start with, imagination does not thereby fail because the noun-concept is malfunctioning in non-I-recruiting PITEs. This is because the architecture of imagination can supplement the shortcomings of the noun-concept. As we saw (Section 4), scripts are unrestrictive in that the script elaborator can tease out details that are neither informed by scripts nor the combination of the UpDater-unfiltered-out beliefs and the imagination premise. If so, then notwithstanding the malfunctioning of the noun-concept, the script elaborator will supply the details needed to ensure the success of imagination in non-I-recruiting PITEs. Put differently, the noun-concept cannot be descriptively rich to such an extent that the script elaborator becomes superfluous. In short, the script elaborator ensures the success of imagination even though the noun-concept is malfunctioning in non-I-recruiting PITEs.

In addition, the details the Wilkes-Van Inwagen-Schectman challenger demands are, contrary to what they say, irrelevant to non-I-recruiting PITEs. Earlier, we saw that a script is generated for an event on account of the event's repeatedness: e.g., by repeatedly engaging in PITEs, a PITE script is generated (Section 4). This, I said, is why Parfit has a fission script and a first-year philosophy student may not. If so, then Parfit could have fleshed out fission with more details than he did—even along the lines Wilkes (1988: 11) enumerates. He presumably did not because such details were irrelevant to the points he wanted to make. Here is why.

First, we do not live in a splitting world, so it is unclear why *sociological* facts about splitting worlds should be important to us. As Kripke (1980) complains similarly about Lewis' (1971) counterpart theoretic framework of possibilia: what our counterparts in possible worlds do is irrelevant to what happens to us in the actual world. Second, we are after metaphysical, not sociological, conclusions, and we can draw them from hypothetical situations that are sociologically under-described. After all, not only do we not live in a world where cats are both dead and alive but the world is also sociologically under-described, yet we infer the relativistic nature of time from such a world. Simply, Parfit is licensed to draw metaphysical conclusions from fission even though it is under-described.

The Wilkes-Van Inwagen-Schectman challenger may say that I have missed the point of their challenge, which is that providing the details would have made imagining the scenario easier. Though some theorists have caved to this line of response—"the details simply go to making the scenario more easily imaginable" (Beck 2016: 124)[7]—I

[7] I am unsure why Beck concedes this point, however. I read him as saying the details the Wilkes-Van Inwagen-Schectman challenge demands are irrelevant to the imagined scenario. His view, which I agree with and discuss below in the main text, is that the challenge mistakes which belief system is integral to imagining the scenario. The Wilkes-Van Inwagen-Schectman challenge thinks it is some non-actual belief system that's actualized for the imagined scenario, whereas what is needed is non-actualizing our actual belief system.

want to dig in my heels. I do not think I have missed the point because I doubt that any of Wilkes, Van Inwagen, and Schectman would agree with this interpretation. What they are saying is rather that once the details they demand are provided, it becomes clear that we are not imagining what we think we are imagining at all, i.e., the details would make imagining the scenario more difficult, not easier.

I contend, however, that they only say this because they wrongly think that the details are relevant to fission, such that the relevance justifies why not providing them is fatal for fission. Having seen that the details are, in fact, irrelevant to fission, it follows that the Wilkes-Van Inwagen-Schectman challenge is unfounded, and so non-I-recruiting PITEs are not descriptively impoverished. They have just the right amount of background details they need, and we are imagining what we think we are imagining with them. This calls to mind Berto and Jago's clarification about how imagination operates: "It's important, however, not to treat agents as importing too much background information into acts of imagination. We do not indiscriminately import arbitrary, unrelated contents into imagined scenarios […] exercises of imagination must obey some constraint of relevance" (2019: 144). Put simply, imagination does not work in the way the Wilkes-Van Inwagen-Schectman challenge wants.

There is more. We saw that one reason the Wilkes-Van Inwagen-Schectman challenge is plausible is that we are supposed to employ our current concepts in the imaginative process even though our noun-concepts have different mental files. Since everyone agrees that different nomological laws hold in possible worlds such that we cannot observationally test the accuracy of our concepts, the Wilkes-Van Inwagen-Schectman challenger would add that we cannot know what we would say, and "what we would say" is the fulcrum on which PITE scenarios turn (Fodor 1964, Ricœur 1992, Wagner 2016). It is unclear to me, however, why what we would say *in* the described possible world matters—as I have said, we do not live there, so why should we worry about some putative belief system that we would hold there? Simply, the issue is not "'What would our beliefs *in the context* be if such-and-such were the case?' [But] 'What do we say *in our context* if such-and-such were the case'" (Beck 2006: 43, original italics). The issue is not actualizing some non-actual belief system for non-I-recruiting PITEs but non-actualizing our actual belief system.

This correction, of course, is backed by the cognitive architecture of imagination. As we have seen, imagination operates solely by manipulating our actual beliefs (Sections 2 and 4). This is why the content of the belief box is copied into the imagination box once the imagination premise is generated by the script elaborator: the agent's actual web of beliefs (occurrent and dispositional) is used as premises during imagination. The belief box only contains actual beliefs. Even scripts, which supply details that are not inferable from our background knowledge, are components of actual beliefs. In short, a scenario is imaginable if and only if the agent has either occurrent (conscious and unconscious) or dispositional beliefs about it. Priest can imagine 1+1=7 because he has at least dispositional beliefs about it (inferring from his paraconsistent logical theoretical assumptions). Whereas because I lack both occurrent and dispositional beliefs about it, given that I am no paraconsistent logician, I cannot imagine it. Unlike Priest, I cannot maintain a coherent reasoning process if I manually hijack my UpDater, telling it to override any belief that would block me from imagining 1+1=7.

Lastly, the Wilkes-Van Inwagen-Schectman challenger may say that even if our actual belief system is at work, non-I-recruiting PITEs cannot show what ought to matter to everyone. That is, since there is no universal belief system that applies to everyone, even if imagination works with the imaginer's actual belief system, only subjective, not objective, normative conclusions can be drawn from it (Martin 1997; Rovane 1997; Baker 2000). This residual challenge does not say that we should not draw metaphysical conclusions from non-I-recruiting PITEs, but that the drawn metaphysical conclusions would lack the dispositive force they *ought* to have because they would only apply to individuals, not everyone. Simply, what follows from non-I-recruiting PITEs is not indicative of what obtains in real life in that the normative conclusions are not factual. It is unclear to me, however, why normative conclusions must be factual.

Why must what "*ought* to matter" matter to everyone? Answer: it must not. It is not a requirement for normative conclusions that they apply to everyone; there is room for disagreements. I may say, "you ought to be friendly with your neighbors," and you may counter, "what if they are nosy and annoying?" In short, normative conclusions, either physical (as with being friendly with your neighbors) or metaphysical (as with PITEs), are contested, so they need not apply to everyone.

But that's not all: the oughtness of normative claims seems to override these disagreements. What I mean is that we often admit differences in what ought to matter to different people, respect their choices, and still say, "even so, what ought to matter to you is so-and-so." Simply, the oughtness of a normative claim overrides whatever differences of opinion there may be among different agents. You may say, "what matters to me when identity does not obtain is numerical identity," and someone else may respond, "that's okay, but what ought to matter to you is psychological continuity." This, in part, is what Parfit aims to demonstrate with fission, which is that regardless of whether you think numerical identity is what matters in the absence of identity, fission shows that what ought to matter to you is psychological continuity.

This view applies to thought experiments even outside philosophy. For example, the Einstein-Bohr disagreement about entangled particles,[8] which asks whether physical reality exists independent of

[8] Quantum entanglement occurs when two or more particles interact in a way

our ability to observe it. Einstein said yes; Bohr said the question is meaningless. We now know, thanks to John Bell some 30 years after the debate, that Einstein was wrong: there are indeed limits on the predicted correlations between entangled particles. The diagnosis of this resolution in cognitive psychological terms is now clear given what I have said in this paper: the interferences in Einstein's script are the fatal kinds á la those in the non-reductionist's fission script.

In conclusion, to the extent to which the metaphysical conclusion drawn from non-I-recruiting PITEs is normative, the cognitive architecture of imagination allows a plurality of them, leaving room for how one can trump another. If so, then Nichols is right: unlike I-recruiting PITEs, there are no dangers to descrying the nature of the self from non-I-recruiting PITEs.[9]

References

Baker, L. R. 2000. *Persons and Bodies: A Constitution View*. Cambridge: Cambridge University Press.

Beck, S. 2006. "These Bizarre Fictions: Thought-Experiments, Our Psychology and Our Selves." *Philosophical Papers* 35 (1): 29–54.

Beck, S. 2016. "Technological Fictions and Personal Identity: On Ricoeur, Schechtman and Analytic Thought Experiments." *Journal of the British Society for Phenomenology* 47 (2): 117–132.

Berto, F., and Jago, M. 2019. *Impossible Worlds*. Oxford: Oxford University Press.

Brown, J. 1986. "Evaluations of Self and Others: Self-enhancement Biases in Social Judgments." *Social Cognition* 4 (4): 353–376.

Chalmers, D. 1996. *The Conscious Mind*. Oxford: Oxford University Press.

Connors, M., and Halligan, P. 2015. "A Cognitive Account of Belief: A Tentative Road Map." *Frontiers in Psychology* 5: https://doi.org/10.3389/fpsyg.2014.01588

Fodor, J. 1964. "On Knowing What We Would Say." *Philosophical Review* 73 (2): 198–212.

Fodor, J. 1987. *Psychosemantics*. Denver: A Bradford Book.

Gendler, T. 2002. "Personal Identity and Thought-Experiments." *Philosophical Quarterly* 52 (206): 34–54.

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

Lewis, D. 1971. "Counterparts of Persons and Their Bodies." *Journal of Philosophy* 68 (7): 203–211.

Martin, R. 1997. *Self-Concern*. Cambridge: Cambridge University Press.

that each particle's quantum state (i.e., the probability distribution for the outcomes of each possible measurement) cannot be described independently of the quantum state of the others, however large the distance between the particles.

[9] Special thanks to Simon Beck for reading an earlier version of this paper and for his useful comments. An anonymous reviewer for this journal also provided significant comments, which helped to solidify the paper's contribution.

Miyazono, K., and Liao, S. 2016. "The Cognitive Architecture of Imaginative Resistance." In A. Kind (ed.). *The Routledge Handbook of Philosophy of Imagination*. New York: Routledge, 233–246.

Nichols, S. 2004. "Imagining and Believing: The Promise of a Single Code." *The Journal of Aesthetics and Art Criticism* 62 (2): 129–139.

Nichols, S. 2006a. "Just the Imagination: Why Imagining Doesn't Behave Like Believing." *Mind and Language* 21 (4): 459–474.

Nichols, S. 2006b. *The Architecture of the Imagination*. Oxford: Clarendon Press.

Nichols, S. 2007. "Imagination and Immortality: Thinking of Me." *Synthese* 159 (2): 215–233.

Nichols, S. 2008. "Imagination and the I." *Mind and Language* 23 (5): 518–535.

Nichols, S., and Stich, S. 2003. *Mindreading*. Oxford: Oxford University Press.

Omoge, M. 2021. "Imagination, Metaphysical Modality, and Modal Psychology." In C. Badura and A. Kind (eds.). *Epistemic Uses of Imagination*. New York: Routledge, 79–99.

Omoge, M. Forthcoming. "On the Place of Imagination in the Architecture of the Mind." In E. Sullivan-Bissett (ed.). *Belief, Imagination, and Delusion*. Oxford: Oxford University Press.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Priest, G. 2016. "Thinking the Impossible." *Philosophical Studies* 173 (10): 2649–2662.

Ricœur, P. 1992. *Oneself as Another*. Chicago: University of Chicago Press.

Rovane, C. 1997. *The Bounds of Agency*. Princeton: Princeton University Press.

Russell, B. 1911. "Knowledge by Acquaintance and Knowledge by Description." *Proceedings of the Aristotelian Society* 11: 108–128.

Schank, R., and Abelson, R. 1977. *Scripts, Plans, Goals, and Understanding*. New York: Psychology Press.

Schechtman, M. 2014. *Staying Alive*. Oxford: Oxford University Press.

Shoemaker, S. 1999. "On David Chalmers's The Conscious Mind." *Philosophy and Phenomenological Research* 59 (2): 439–444.

Taylor, S., and Brown, J. 1988. "Illusion and Well-being: A Social Psychological Perspective on Mental Health." *Psychological Bulletin* 103 (2): 193–210.

Unger, P. 1990. *Identity, Consciousness, and Value*. Oxford: Oxford University Press.

Van Inwagen, P. 1997. "Materialism and the Psychological-Continuity Account of Personal Identity." *Philosophical Perspectives* 11: 305–319.

Wagner, N. 2016. "Transplanting Brains?" *South African Journal of Philosophy* 35 (1): 18–27.

Wilkes, K. 1988. *Real People*. Oxford: Oxford University Press.

Williams, B. 1970. "The Self and the Future." *The Philosophical Review* 79 (2): 161–180.

Williams, B. 1973. "Imagination and the Self." In B. Williams. *Problems of the Self*. Cambridge: Cambridge University Press, 26–45.

Wilson, T., and Brekke, N. 1994. “Mental Contamination and Mental Correction: Unwanted Influences on Judgments and Evaluations.” *Psychological Bulletin* 116 (1): 117–142.

Wilson, T., Centerbar, D., and Brekke, N. 2002. “Mental Contamination and the Debiasing Problem. ” In D. Griffin, D. Kahneman, and T. Gilovich (eds.). *Heuristics and Biases*. Cambridge: Cambridge University Press, 185–200.

Croatian Journal of Philosophy
Vol. XXIII, No. 67, 2023
https://doi.org/10.52685/cjp.23.67.5
Received: December 29, 2021
Accepted: October 5, 2022

Is Autism a Mental Disorder According to the Harmful Dysfunction View?

MLADEN BOŠNJAK
University of Rijeka, Rijeka, Croatia

The supporters of the neurodiversity movement contend that autism is not a mental disorder, but rather a natural human variation. In a recent paper Jerome Wakefield, David Wasserman and Jordan Conrad (2020) argued against this view relying on Wakefield's harmful dysfunction theory of mental disorder (the HD theory). Although I argue that the HD theory is problematic, I contend that arguments offered by Wakefield et al. (2020) against those of the neurodiversity movement are plausible, except in one respect: their claim that high functioning autism in general is not a disorder is not well supported. I argue instead that the disorder status of high-functioning autistic persons should be judged on a case-by-case basis, depending on the harmfulness of the condition. In this regard, I maintain that the list of basic psychological capacities provided by George Graham (2010) provides an adequate conceptualization of harm. Moreover, I show how this framework may offer an appropriate tool for a case-by-case assessment of harm associated with high-functioning autism.

Keywords: Autism; neurodiversity movement; social model of disability; harmful dysfunction; harm in psychiatry; basic psychological capacities.

1. *Introduction*

Since Leo Kanner (1943) introduced the notion, there have been many controversies around it, including whether it is a mental disorder (Wakefield, Wasserman, and Conrad 2020). The disorder status of autism is relevant for determining treatment and other appropriate social responses to the condition like, for instance, the criminal responsibility of autistic offenders (Bošnjak 2022, Malatesti, Jurjako and Meynen 2020). While the medical view is that autism is a mental disorder (APA 2013, Cushing 2018), proponents of the neurodiversity movement disagree (Blume 1998, Meyerding 2014, Sinclair 1993, Armstrong 2015, Chapman 2019, Jaarsma and Welin 2012, Ortega 2009, for discussion see Hughes 2021). Jerome Wakefield, David Wasserman, and Jordan Conrad (2020) have recently made progress on this issue by discussing it in the context of an account of mental disorder. This is Wakefield's influential harmful dysfunction analysis of mental disorder (HD for short) (see Wakefield 1992, 2007, 2014).

The aim of this paper is to discuss Wakefield et al.'s (2020) criticism of the arguments advanced by the advocates of the neurodiversity movement who deny that autism is a mental disorder. Although I do not subscribe to all aspects of Wakefield´s HD account of mental disorder, I agree with Wakefield et al.'s (2020) rebuttals of the arguments offered by the proponents of the neurodiversity movement. However, I question their claim that high functioning autism is most likely not a disorder. I argue that a general conclusion on this matter cannot be decided in advance for all cases. Rather, it should be decided on a case-by-case basis depending on how and in what way high-functioning autistics are harmed by their condition, if they are harmed by it at all. However, the HD view does not offer a helpful account of harm to adjudicate this question. To make progress on this problem, I argue that the list of basic psychological capacities offered by George Graham (2010) provides an appropriate elaboration of the concept of harm and a useful framework for such a case-by-case assessment of harm that is relevant for mental disorder.

In the paper, I proceed as follows. I first present the conceptualization of autism spectrum disorder as depicted in the fifth edition of the *Diagnostic Statistical Manual* (from now on DSM-5, American Psychiatric Association, APA, 2013). Then I move on to present the claims of the supporters of the neurodiversity movement. I contend that a proper evaluation of their arguments should be based on the backdrop of a general account of mental disorder. I argue that the evaluation of these arguments, offered by Wakefield et al. (2020) is convincing. Nonetheless, I criticize Wakefield´s account of mental disorder (1992, 2007, and 2014) and opt for a more general hybrid account of disorder that does not rely on a specific notion of dysfunction. Finally, I rely on the list of basic psychological capacities offered by George Graham to address the issue of the disorder status of high-functioning autism.

2. *Autism in the DSM-5*

According to the DSM-5, autism spectrum disorder is a neurodevelopmental disorder characterized by a lack of empathy, a deficit in verbal and nonverbal communication, difficulties in understanding and maintaining human relationships, having a limited range of interests, repetitive behaviors, and problems in adjusting behavior to different circumstances (APA 2013: 299.00; F84.0).

Symptoms are divided into two categories: (1) Social Communication and (2) Restricted and Repetitive Behaviors. The DSM-5 differentiates three levels of symptom severity: level 1 ("Requiring support"), level 2 ("Requiring substantial support") and level 3 ("Requiring very substantial support"). Level 1 includes autistics who live independently and have a satisfactory quality of life despite problems in social communication and struggles in adapting to changes, starting and maintaining conversation, and having lower interest in social interaction. These obstacles require behavioral therapy. Level 2 encompasses autistics with social impairments, decreased verbal and nonverbal communication abilities and slight behavioral inflexibility (e.g., difficulties in dealing with changes, limited interest, and lower reactivity to social cues). They need assistance and therapy to achieve a good quality of life. Level 3 covers autistics with minimal social interactions, who mostly lack the ability to speak. They have significant problems in everyday functioning and adapting to environmental changes.

In the previous edition of the DSM, the terms *Asperger syndrome,* and *Pervasive Developmental Disorders—Not Otherwise Specified* were used to mark autism of level 1, *Rett syndrome* and *Childhood disintegrative disorders* to mark level 3 autism (APA 1994: DSM IV). In the newest edition, these categories were placed on a single spectrum. Thus, autism is a heterogeneous disorder, including people with severe learning and verbal impairments as well as high-functioning individuals with a potentially outstanding IQ (Feather 2016).

From the medical perspective described in the DSM-5, autism is a mental disorder. Mental disorder in the DSM-5 is defined as follows:

> A mental disorder is a syndrome characterized by clinically significant disturbance in an individual's cognition, emotion regulation, or behavior that reflects a dysfunction in the psychological, biological, or developmental processes underlying mental functioning. Mental disorders are usually associated with significant distress or disability in social, occupational, or other important activities. An expectable or culturally approved response to a common stressor or loss, such as the death of a loved one, is not a mental disorder. Socially deviant behavior (e.g., political, religious, or sexual) and conflicts that are primarily between the individual and society are not mental disorders unless the deviance or conflict results from a dysfunction in the individual, as described above. (DSM 5: 20).

Autism satisfies the above definition of mental disorder because it typically involves "disability in social, occupational, or other important activities" which appear in early developmental period and are thought to be caused by some kind of neurobiological dysfunction (for an overview of the dominant theories of autism, see Fletcher-Watson and Happe 2019). Since autism is included in DSM and it satisfies the above definition of mental disorder, the default position among the medical practitioners seems to be that autism is a mental disorder.

However, self-advocate autistics (both within and outside academia) (Blume 1998, Meyerding 1998, Sinclair 1993, Chapman 2019) and other academics (Armstrong 2015, Jaarsma and Welin 2012, Ortega 2009) argue that autism is a normal human variation in brain functioning. Thus, the proponents of the neurodiversity movement claim that autism is not a mental disorder, or that at least some of the autistics on the spectrum should not be considered as having a disorder. Arguments of such type usually presuppose a specific view about what it means to be disabled in everyday functioning. So, in the next section I provide a short overview of the disability theory which is relevant for understanding the arguments advanced by the neurodiversity movement supporters.

3. *The neurodiversity movement against the medicalization of autism*

Many of the claims endorsed by the neurodiversity movement are often based on the backdrop of a family of views that fall under the *social model of disability*. In what follows, I provide an overview of the main claims underlining this model.

Many influential publications on disability distinguish between impairment and disability. On the one hand, impairments are seen as "problems in body function or structure such as a significant deviation or loss" (World Health Organization 2001: 10). On the other hand, in various documents such as the *International Classification of Functioning, Disability and Health,* the U.N. *Standard Rules on the Equalization of Opportunities for People with Disabilities*, the *Disability Discrimination Act* (U.K.), and the *Americans with Disabilities Act* (U.S.), disability is construed as "(1) a physical or mental characteristic labeled or perceived as an impairment or dysfunction … and (2) some personal or social limitation associated with that impairment" (Wasserman et al. 2016).

There are two principal perspectives on disability: the medical and the social model. According to the medical model, the physical or mental incapacities of people cause the barriers that limit their daily functioning. In contrast, the social model emphasizes society's role in limiting the daily functioning of people considered as disabled. Thus, the focus, instead of being on the characteristics of the person as in the medical model, is on the inappropriate environment and social organization (Wasserman et al 2016). For example, it is not the bodily or physical impairments which render most buildings in the city of Rijeka inaccessible for wheelchair users, but the absence of ramps and elevators.

Some of the claims made by the supporters of the neurodiversity movement are also related to claims made by supporters of movements for civil rights, such as the movement for LGBT rights, as well as with the antipsychiatry movement. Both the neurodiversity and antipsychiatry movement agree that psychiatry is often used as a means of oppressing people whose behavior does not fit with the prevailing social norms and values. Some also argue that severe autism should be treated and thus considered a mental disorder (for more about this topic, see Graby 2015). However, the proponents of the neurodiversity movement argue that the need for specific resources for autistics does not imply that autism should be considered a mental disorder (Nicolaidis 2012, Den Houting 2019, Legault et al. 2019, Legault et al. 2021). In other words, promoting an ideal of social justice and change in policies and arguing for a more adequate view of autism as non-disorder are not mutually exclusive. It is important to keep in mind that one of the main aims of the neurodiversity movement is to combat stigma. This motivates the most radical proponents of the neurodiversity movement to even deny the disorder status to the whole autism spectrum. I think that the aims of the neurodiversity movement such as destigmatization and equal rights of autistics persons are very desirable. Nonetheless, I think that the denial of the disorder status is not the right approach to achieve these goals. I strongly believe that it is consistent to claim that autism is a mental disorder and at the same time to demand equal rights and to fight against stigma. Moreover, I think that philosophers can offer theoretical frameworks and arguments for reconciliation between the medical perspective on autism and the neurodiversity movement (see, e.g. Nelson 2021)

However, before examining the claims of the supporters of the neurodiversity movement, we need a general framework within which we might evaluate them. Relevant for our context is a framework that can help us to decide whether a condition is a mental disorder. Thus, in what follows, I turn to this issue.

4. *A harmful dysfunction account of mental disorder*

A useful way to approach this issue is offered by Jerome Wakefield, David Wasserman and Jordan Conrad (2020). They presuppose Wakefield's influential account[1] of mental disorder (e.g., 1992, 2007, and 2014). The core of the account is summarized in the following quote:

> A condition is a disorder if and only if (a) the condition causes some harm or deprivation of benefit to the person as judged by the standards of the person's culture (the value criterion), and (b) the condition results from the inability of some internal mechanism to perform its natural function, wherein natural function is an effect that is part of the evolutionary explanation of the existence and structure of the mechanism (the explanatory criterion). (Wakefield 1992: 384)

Wakefield relies on an etiological theory of natural function (see, e.g. Šustar and Brzović 2014). According to this theory, natural func-

[1] For an overview of theories of mental disorder, see, e.g. Cooper (2007, ch. 3) and Bolton (2006).

tion of some system is determined by its evolutionary history, i.e., by natural selection, which "designed" the system to perform a particular function. For example, we can ascertain that the function of the heart is to pump blood because organisms that had organs with such a capacity during evolutionary history outlived and left more offspring than their conspecifics.

Wakefield thinks that if a condition is a mental disorder then it must be both harmful and caused by a dysfunctional physical or psychological mechanism. The following two examples illustrate these two components. Even if there was a dysfunction in the case of homosexuality, this condition is not a disorder because it is not by itself harmful. If homosexuality is associated with harmful consequences, then this harm would be extrinsic, most likely caused by negative and stigmatizing attitudes of the other members of the society. Alternatively, in the case of antisocial personality disorder (ASPD), a person with ASPD is harmed because their behavior often gets them into trouble for which they spend much time in prison. However, such a condition would not be a disorder unless it is underpinned by a psychological or biological dysfunction (Jurjako 2019).

There are several reasons for adopting something akin to Wakefield's hybrid or two-component account of mental disorder. First, Wakefield's account is extremely influential and has been used to discuss and adjudicate the disorder status of many conditions and symptoms, including delusion (e.g. Lancellotta and Bortolotti 2020), misbelief (e.g. McKay and Dennett 2009) and psychopathy (e.g. Jurjako 2019). Second, in broad strokes, Wakefield's account nicely fits with how mental disorder is conceptualized in the dominant psychiatric diagnostic manuals, such as e.g., DSM and ICD (Murphy 2006: 35; Biturajac and Jurjako 2022; cf. Amoretti and Lalumera 2019). The third reason is its explicit inclusion of the notion of harm, which I take to be indispensable for thinking about the nature of disorder (see, also, Biturajac and Jurjako 2022). I maintain that the key role of medicine (but not the only one) is to cure or treat disorders. But if some condition is not harmful, there is, *prima facie*, no reason to cure or treat it, and, thus to think of it as a disorder. Of course, we often medically treat conditions that are not disorders, such as pregnancy. Nonetheless, we can all agree that even in such cases, the default presupposition is that there is no medical reason to treat a condition if there is no dysfunction that might actually or potentially harm a person.

Despite the positive sides of Wakefield's HD account, it still relies on some controversial assumptions. In fact, both the dysfunction and harm aspects of HD have been extensively criticized (see, e.g. McNally 2001; Bolton 2006; Murphy 2006; Bingham and Banner 2014; Murphy-Hollies 2021). More specifically, some argue that there could be disorders whose causal basis is a consequence of adaptation (see, e.g. Garson 2021). For example, a person who has been raised in an abusive environment might develop antisocial personality traits as a developmental adaptation to such environment. Moreover, these traits might still be adaptive if the person continues to live in uncertain, violent, and otherwise difficult circumstances. However, if such a person would be transferred to a nonviolent and friendly environment, then antisocial traits would fail to be adaptive because they would likely lead to frequent incarceration which as a consequence would cause an inability to perform normal social and occupational activities, reduction in well-being, and it would have other harmful effects (for discussion, see Jurjako 2019). This example illustrates that traits comprising a condition could be adaptive and thus functional, but still associated with a disorder. Moreover, a more general problem for relying on an etiological reading of the dysfunction component is that it is not clear whether it would be possible to practice medicine until the evolutionary role of different mechanisms and organs is discovered (see, e.g. Bolton 2006). The problem is that if we accept Wakefield's theory of mental disorder, we would not be able to determine the disorder status of many conditions that are thought to be disorders. Namely, it seems practically impossible to reliably establish whether or not some condition is caused by a failure of some mechanism to perform its evolutionary designed function because the evidence about evolutionary past of such mechanisms is not available to us, and most likely will never be.

The problem with Wakefield's view of harm is its cultural relativity and underspecificity. Wakefield (1992) typically construes harm as something that is negatively judged by our society without providing additional criteria how this might be determined (see also Wakefield and Conrad 2019). This view makes the mental disorder status relative to sociocultural standards adopted by a particular society. In this regard, Rachel Cooper (2021: 537) notes that Wakefield´s concept of harm falls short "because whole societies can be wrong in how they evaluate a condition". In addition, it is plausible to think that there are conditions, such as schizophrenia, that are associated with low quality of life, often leading to fatal outcomes, and as such can be considered as harmful regardless of the evaluative standards entrenched in a specific society in which it occurs. Moreover, even if we leave the problem of cultural relativism aside, Wakefield does not really offer a substantive view of harm that can be used to adjudicate difficult cases (see, also Cooper 2021: 538). This issue will become important once I discuss the disorder status of high-functioning autistics. I will argue that assessments of harm in the case of high-functioning autism will not be solved if we do not adopt a more concrete account of harm. Wakefield's view of harm as something that is negatively judged by our society is too vague to perform this task. To remedy this problem, in section 6, I will argue that we should adopt the list of basic psychological capacities offered by George Graham as a useful way to conceptualize harm and estimate it in the case of high-functioning autism.

For the foregoing reasons I do not accept Wakefield´s harmful dysfunction account of mental disorder in its entirety. Nonetheless, for the purposes of this paper I adopt it insofar it evinces a hybrid view of mental disorder. In general, hybrid views presuppose that disorders have causal basis that produce harmful effects that can or should be medically treated (see e.g. Stegenga 2015, Biturajac and Jurjako 2022). For the present discussion it is not important whether such causal bases will be interpreted in terms of an etiological theory of dysfunction or some other view. The important thing is that however we understand the dysfunction part of the disorder, it should be associated with significant harmful effects.

Presupposing such a hybrid view of mental disorder, in the next section, I will provide an overview of Wakefield et al.'s (2020) discussion whether autism is a mental disorder.

5. *The harmful dysfunction view and neurodiversity*

5.1. *The essence of autism and harm*

Neurodiversity advocates often claim that autism does not involve any dysfunction that would warrant the disorder status. One interesting argument in this respect is offered by the clinical psychologist Simon Baron-Cohen who contends that while the "autistic essence" confers many advantages, many of the harms usually associated with autism are not part of the condition. The claim is that whatever harms might be associated with autism, they are only contingently associated with it. Thus, autism *per se* should not be regarded as a harmful condition that is underpinned by dysfunctions. Baron-Cohen offers this kind of argument in the following:

> Some will object that a child with autism who has epilepsy is not an example of neurodiversity but rather he or she has a disorder. And they are right. Epilepsy is a sign of brain dysfunction and causes disorder (fits) and should be medically treated. But epilepsy, while commonly co-occurring with autism, is not autism itself. Others may say that a child who has language delay or severe learning difficulties is not an example of neurodiversity but has a disorder, and I would support their demand for treatments to maximize the child's potential in both language and learning. But again, although commonly co-occurring these are not autism itself. (Baron-Cohen 2017: 744)

In response to this, Wakefield et al. (2020: 507) note that the idea of autism including an essence does not take seriously enough the heterogeneity of autism. Indeed, in contrast to the essentialist perspective, Daniel Weiskopf indicates that autism is more properly construed as "a network category defined by a set of idealized exemplars linked by multiple levels of theoretically significant properties" (2017: 175). Thus, autism as a category is not coherent enough to be considered as "an adaptive trait or a distinct perceptual and cognitive style" that would make plausible the claim that autistics have a shared essence which is distinct from the accompanying physical, psychological or social impairments (Wakefield et al. 2020: 507).

5.2. *Context insensitivity and harm*

Another argument used by the members of the neurodiversity movement is based on Uta Frith's "weak coherence" theory (Frith 1989). According to this account, autistics have a diminished capacity to incorporate data into a coherent whole. Autistics are often preoccupied with details but misunderstand relations between them and their contextual meaning. For example, an autistic person could remember all the details of a story without understanding the meaning of the whole story (Frith 1989, Happė 1999). Interestingly, Frith thinks this might be perceived as an exceptional ability to operate with local data, rather than a handicap. Similarly, advocates of the neurodiversity movement see poor sensitivity towards meaningful context as resulting from a natural biological variation (Baron-Cohen 2009). In addition, it has been discovered that autistics perform better on some cognitive tasks than neurotypicals. For instance, in some situations, unlike the neurotypicals, autistics are immune to optical illusions due to reduced context sensitivity (Doherty et al. 2010).

In response, Wakefield et al. (2020) indicate that context insensitivity is often harmful, and people normally grow out of it. For instance, children with underdeveloped perceptual abilities are also more immune to optical illusions, indicating that people may be less prone to optical illusions once their perceptual capacities mature. Here it is important to note that sensitivity to context seems to be a necessary component of psychological maturing because reduced sensitivity to contextual cues can be life-threatening. Wakefield et al. (2020) illustrate this with the case of an autistic young adult who, while on a cruise ship, jumped overboard because he wanted to take a swim (McLaughlin and Sutton 2018). This behavior might be explained by the context insensitivity which is responsible for an inability to understand the situation and therefore to prevent the impulse to take the swim. From this it can be concluded that context insensitivity can be biologically more harmful than beneficial when it comes to autistic traits.

Moreover, it should be noted that there is a relation between the level of functioning in everyday life activities and impairments in context sensitivity. The level of functioning and impairments in context sensitivity are inversely proportional which means that more severe impairments in context sensitivity imply lower level of functioning and *vice versa*. If there were a balance between the lack of contextual understanding and functioning, then autism could be considered as a beneficial natural variation. However, Wakefield et al. (2020: 509) note that there are many open empirical issues surrounding this claim. In particular, it is undecided whether lower context sensitivity is distinctive of autism or a natural variation in the general population, and whether autistics possess some other capacities which might render lower context sensitivity beneficial.

5.3. *Autism and savant abilities*

There are autistics with special capacities often referred to as “savant abilities”, such as outstanding memory of some types of events, calendrical calculation, precise drawing, and so on, which, according to some researchers, seem to be an integral part of the autistic condition (Happe 2018; Meilleur, Jelenic, and Mottron 2014). In other words, it is not possible to have these capacities without being autistic. This is the reason why some proponents of the neurodiversity movement think of autism as a special but natural way of brain-functioning. However, Wakefield et al. (2020: 510) argue that the savant abilities argument is unpersuasive because in most cases harm caused by autism is more severe than the benefits brought about by savant abilities. The fact that some argue that savant abilities are integral part of autism is the reason why we compare harms associated with autism with benefits stemming from savant abilities. Different disorders can bring about some advantages as well. For example, albinism might be beneficial in environments where there is not much sunlight available because it would allow vitamin D to be synthetized from limited amount of sunlight (Reznek 1987: 86). However, possible benefits of albinism do not *ipso facto* imply that it is not a disorder. In fact, even if it would have such benefits, still we would have reason to think of it as a disorder because people who have it are not protected from solar rays and therefore often suffer from sunburn, have greater chances to get skin cancer, etc.

Moreover, Wakefield et al. (2020: 510) indicate that there are three reasons why having savant abilities does not imply that autism involves a natural variation in brain functioning that is not harmful. First, it is not true that savant abilities are integral part of autism because only 10–25 percent of autistics exhibit savant talents and skills (Happé 2018; Meilleur et al. 2014). Second, collaboration and social interaction are needed to put in effect these capacities, which is not possible in the case of severe autism. Third, savantism can be related to different brain illnesses and brain damage such as frontotemporal dementia (Miller et al. 1998; Treffert 2009). Therefore, it is not true that savantism is distinctive for autism, and because of that autism cannot be considered as a natural variation in brain functioning.

5.4. *Autism as personal identity and culture*

Another type of argument provided by the neurodiversity movement is to suggest that autism is essential to autistics’ personal identities because it confers special mental capacities and a specific world comprehension. Since autism affects the mental life of a person (her beliefs, wishes, and emotions) and mental life is considered to be a crucial part of personal identity, some autistics conceive autism as essential for their personal identity, in contrast to physical disability which is usually not deemed as intimately connected to personal identity.

However, as noted well by Wakefield et al. (2020: 512), the identity possessed by autistic individuals has nothing to do with the question of whether autism is a disorder or not. Thus, even if it is accepted that autism is a crucial part of someone's identity, this would not change the fact that autism might also be a disorder.

According to some authors autism is a socially constructed category given the heterogeneity and great expansion of it in DSM through time (see, e.g. Chapman 2016; Cushing 2018). Some authors go as far as claiming that instead of alleged autistic essence, what autistics have in common are properties which have arisen in response to being stigmatized as autistics, which for them means that it makes more sense to view autism as a form of culture rather than a disorder (see, e.g. Sarrett 2016; Verhöff 2012). Moreover, some argue that such autistic communities and culture should be appreciated and maintained (see, e.g. Straus 2013).

To this argument Wakefield et al. (2020: 512) provide a plausible retort. Although a society can influence the formation of autism as a category, this does not tell us anything about whether autism is caused by a dysfunction or whether it is harmful. Furthermore, they assert that the existence of autistic communities has nothing to do with the illness status, since there are many communities of people who share political and religious beliefs, taste in music and movies, dietary habits, and so on. The fact that people who share autistic traits have decided to establish a community does not imply anything about the disorder status of autism and whether it should be treated.[2] The idea that some condition is a disorder which should be treated is fully consistent with having a respect for a community that is based on this disorder. This can be seen in the case of communities of individuals afflicted by different major illnesses, which seek medical treatment of their condition, regardless of the fact that treatment may decrease the amount of community members. Wakefield et al. (2020: 512–513) note that the possibility of extinction is not distinctive only for communities which rest on disorders, but also for the communities which are based on natural diversity among people, such as Western European monastic culture or Yiddish culture in the United States, that disappeared because of assimilation. It is possible to appreciate the decision made by people who accepted the dominant culture while, at the same time, feel remorse because of their cultural extinction, which followed the assimilation. Wakefield et al. (2020: 513) argue that the same thing might happen with the deaf community. This is a hypothetical situation. Imagine that there is a cure for deafness and that deaf people widely welcome it, which consequently leads to the extinction of the deaf community. In this case, we most likely would not see anything intrinsically morally problematic

[2] As I have explained earlier, I think that the default position is that the justification of treatment and disorder status are related. If some condition should be medically treated, the default presupposition is that this condition is a disorder.

about it, because deaf people have freely decided to accept the cure offered to them. Here it is not morally significant whether a community of autistics or deaf people will really go extinct or not. The relevant question instead is whether the potential extinction of these communities would be caused by a decision of their members to accept successful treatment of their condition. So, it does not seem implausible to hold at the same time that autism is a disorder and that, as long as autistic persons give their consent to be treated, there is not an intrinsic moral reason against offering treatment that might undermine the existence of their community.

5.5. *Autism, harm, and a hostile society*

Finally, the most radical proponents of the neurodiversity movement argue that autism is not harmful at all. Such an approach argues that capacities of autistics should be taken as a starting point when assessing their well-being (Robeyns 2016). According to this argument, many cases of autism would not be regarded as harmful if harm is assessed in accordance with the capabilities that autistics actually possess. However, it is obvious that this approach does not work in cases of severe autism. For instance, Wakefield et al. (2020: 513) convincingly indicate that the inability to communicate and form an emotional attachment to others and feeling of sensory overload in public places can seriously impede well-being, however it is conceived.

The proponents of the neurodiversity movement argue that most harms associated with autism are caused by unfriendly environments, which are designed for people with typical brain functioning, similarly to how people with physical impairments are excluded from a society because social environments are designed for people without physical impairments (Jaarsama and Welin 2012, Chapman 2019). Here the claim is that harms suffered by autistics are not a consequence of autism as such. They are, rather, consequences of prejudice and stigmatization and the organization of the social environment or even physical space.

The same sort of argument was applied to the case of homosexuality when it was removed from DSM-III's list of disorders (Jaarsama and Welin 2012, see also Stegenga 2021). There is a distinction between harms caused immediately by a dysfunction and harms that result from a reaction of a society to the condition. This distinction was introduced by Robert Spitzer, who played a key role in de-pathologizing homosexuality in DSM. Together with Paul Wilson, they put forward the definition of disorder as a condition that is "regularly and intrinsically associated with subjective distress" or "impairment" which means that "the source of the distress or impairment in functioning must be the condition itself and not with the manner in which society reacts to the condition" (Spitzer and Wilson 1975: 829, see, also Spitzer and Endicott 1978: 18).

In the case of homosexuality, it is obvious that harm is caused by misconceptions and inappropriate reactions from other members of the society. Proponents of the neurodiversity movement argue that, in the same way, the harms associated with autism are at least partially caused by misconceptions about autism and absence of adjustment (Dominus 2019).

There are two difficulties with such application of the social model to autism (see Wakefield et al. 2020: 514). The first problem is the misuse of the difference between direct/indirect or intrinsic/extrinsic harms. Intrinsic or direct harm is harm caused by the condition itself, while indirect/extrinsic harm is harm caused by unjustifiable stigmatization and prejudices of the society. There are disorders which are related to social interaction, but are nevertheless disorders. Take, for instance, aphasia that is caused by brain trauma. Aphasia is an inability to linguistically communicate that causes problems for social interactions with other people. Thus, harm associated with aphasia can be considered as intrinsic because it will be present regardless of how a society treats people with aphasia. In the same way, harm resulting from autism is socially related, but still it can underpin the disorder status because it is caused by a dysfunction in psychological mechanisms underlying their ability to read mental states of others (Baron-Cohen 1995), lack of the capacity to recognize the influence of their behavior on others (Attwood 1998, Mercier et al. 2000), and difficulties with understanding emotions (Burgoine and Wing 1983, cited in Attwood 1998). It is clear that harms which result from these incapacities have nothing to do with the stigmatization and prejudices of the society toward autistics, although they are socially related. Since these incapacities are intrinsically associated with autism and they cause harm to them it is very likely that even changes in social practices would not help to significantly reduce harm. Thus, we have reason to think that the disorder status of autism is warranted (see Wakefield et al. 2020: 515).

Although autism seems to be an intrinsically harmful condition, still we might ask what a society can and should do to ameliorate the level of social detriments experienced by autistic persons. It is plausible to think that the magnitude of harm suffered by autistics is also influenced by external factors, such as the perception of autism in a society and the way the society treats autistics. We can also agree that this influence is higher than in the case of, for instance, aphasia. Were it to be the case that the social price of decreasing negative impacts of autism is low, it would be sensible to expect a society to adjust to the needs of autistics. However, it is not immediately clear when this will be the case.

Chong-Ming Lim (2017) indicates several things that should be considered when assessing whether the adjustments are sensible or not, such as finances and demands for neurotypicals to change their behavior, fundamental conventions, and values. With respect to this, I think that Wakefield et al. (2020: 514) correctly conclude that it is not sensible to demand from neurotypicals to change their social conventions regarding paying attention to emotional cues, contexts, and conversational implicatures. Although such a change in social conventions would be beneficial for autistics, it is clear that it would not be feasible to introduce it for the rest of the population.

Wakefield et al. (2020: 515) argue that the second problem regarding attempts to reconcile autism with the social model of disability is the heterogeneity of autistic conditions. It is plausible that only high-functioning autism fits well with the social model because disabilities associated with many cases of high-functioning autism could be successfully reduced by environmental and social adjustments in contrast to typical cases of severe autism.

Thus, Wakefield et al. (2020: 504) contend that moderate neurodiversity is a plausible position. Moderate neurodiversity acknowledges the disorder status of classic severe autism but doesn't qualify as disorders high-functioning autism and what was formerly entitled Asperger's syndrome. This position is in-between strong neurodiversity, which is the claim that the whole autism spectrum is not a mental disorder, and weak neurodiversity which claims that the present classification of autism should remain unchanged.

I agree with Wakefield, Wasserman, and Conrad (2020) that the reviewed arguments of the neurodiversity advocates are not plausible, but I disagree with their view that high-functioning autism is likely not a mental disorder. As a class, high-functioning autism is also very heterogeneous (Weiskopf 2017). For this reason, we cannot give one ultimate answer to the question whether high-functioning autism is a disorder or not. Any general claim on this matter would be inappropriate, both because of our present lack of knowledge and conceptual issues regarding the distinction between high-functioning autism and low-functioning autism. There are no clear criteria on how to precisely distinguish between these two categories and as Wakefield et al. (2020: 505) notes "we should expect disagreement and uncertainty in many cases".

I think that some cases of high-functioning autism can be thought of as involving a disorder, while other cases should not be thought of as involving a disorder. Because of this I think that in each case individual assessment of functioning should be made. In other words, we should assess whether the cognitive and social impairments typically associated with high-functioning autism are such that they cause sufficient harm to autistic individuals. However, as mentioned earlier, what is needed to solve this issue is a more elaborated concept of harm than the one offered by Wakefield (1992). To start solving this problem we should have a working account of what are the relevant cognitive and social abilities which are needed for everyday normal functioning and how their impairment might be harmful to high-functioning autistics. Thus, in the next section I argue for what I believe to be a good further elaboration of the relevant capacities that will provide a valuable tool for assessment of harm in cases of high-functioning autism. Given the limited space, this account can only be provided in broad outlines, but, still, it should be informative enough for showing how we can use it for determining in individual cases whether high-functioning autistics should be considered as mentally ill.

6. *Capacities, harm, and high-functioning autism*

We can all agree that some condition is harmful to a person if it significantly interferes with her well-being and functioning. However, to adjudicate whether a high-functioning autistic person is harmed by some condition in a way that is relevant for determining whether they suffer from a mental disorder, we need to be able to determine the relevant forms of harm and their causal bases. I maintain that this question may be approached by thinking about the psychological capacities that are necessary for leading a healthy and satisfying life. Earlier we saw that a plausible view of mental disorder requires that harm should be intrinsic, in the sense that harm is caused by an internal impairment in a relevant psychological capacity and not by stigmatization or prejudice. Moreover, the impairments in fundamental capacities that cause harm need to be such that they most likely cannot be ameliorated by introducing changes in social practices or environment. If a condition is harmful and a consequence of an impairment in the relevant psychological capacity, then we would have reason to think of this condition as a mental disorder. Now the pertinent question is what are these psychological capacities which are necessary for leading a healthy life?

I maintain that the list of basic psychological capacities offered by George Graham (2010: 147–148) provides a particularly good elaboration of what is relevant for assessing the kind of harm that underpins mental disorders. Graham claims that his list provides basic psychological capacities because they pass the veil of ignorance test as formulated by John Rawls (1971). Rawls uses the veil of ignorance to illustrate a hypothetical situation in which free, equal, and rational agents choose basic principles of justice, without knowing anything about their gender, race, nationality, and socioeconomic status. Analogously, Graham (2010, 139–142) uses this model to determine the list of basic psychological capacities that are universally needed for a decent life by all people, regardless of their specific condition. Graham contends that by thinking about this issue from the perspective of a veil of ignorance, where a person tries to decide what are the capacities that “no one (…) would wish to be without or to have seriously compromised or impaired” (Graham 2010: 154), we will come to see the following list of capacities as fundamental: 1) Bodily/spatial self-location, 2) Historical/temporal self-location, 3) General self/world comprehension, 4) Communication, 5) Care, commitment and emotional engagement, 6) Responsibility for self and 7) Recognition of opportunities or "affordances" (Graham 2010: 147–149).

In what follows, I will summarize Graham's descriptions of the capacities that pass the test of veil of ignorance and are relevant for the discussion of harm in the case of high-functioning autism.

1) Communication. To be able to communicate with each other about ourselves and the world, we must possess sufficient listening and speaking competencies in some system of communication (e.g., one's mother tongue, sign language, etc.). In interactions with others, we assess the soundness of others' utterances, but to do this, we first need to understand their meaning. Communication is an important source of information, and it connects people with each other (Graham 2010: 148).
2) Care, commitment and emotional engagement. People are usually committed to and take care of things and people they consider important and as a consequence, they feel bad if things or people they care about are in some way endangered, or feel happy if they are not (Graham 2010: 148–149).
3) Responsibility for self. We are able to take care for ourselves, which means that we can control our behavior by forming intentions, assessing the impulses and inhibitions, making practical decisions and self-reflective choices. We can conform our behavior to our decisions and choices; mostly we do not behave impulsively (Graham 2010: 149).
4) Recognition of opportunities or "affordances". We are able to recognize different possible choices we can make in the process of decision-making. Although many people feel great deal of anxiety about making decisions, people usually want to make autonomous decisions in life, which presupposes being aware of different paths and opportunities available to them (Graham 2010: 149).

Using Graham's account of psychological capacities for assessing the mental disorder status of high-functioning autistics is appropriate because it satisfies two important desiderata. First, this account is unique in analytic philosophy of psychiatry in that it provides a concrete list of psychological capacities that is specifically made for testing particular cases of mental disorder. Second, the list of capacities is justified via an ethical procedure (i.e. the veil of ignorance) that purports to be fair and provide universal standards that can be accepted across different cultures. Thus, the justification of these capacities is not vulnerable to unjustified forms of cultural relativism, because they are "not derived from our individually variable desires or capacities, but from competencies that we are bound to value and need, regardless of which specific goals we possess and pursue" (Graham 2010: 147).

Following the symptomatology of autism from DSM-5 (see above section 2), we can plausibly say that capacities of communication and emotional engagement are often impaired even when it comes to high-functioning autistics. I think that capacities underpinning responsibility for self and recognition of opportunities might also be impaired because even high-functioning autistics show repetitive behavior, and they possess a limited range of interests. Moreover, repetitive behavior might be caused by an inability to control impulses and inhibitions. Finally, autistics have difficulties recognizing the needs and mental states of other people that can be expressed by various social cues, such as facial expressions and tone of voice. This likely leads to impairments in functioning in everyday social interactions.

However, whether these impairments cause harm that would trigger the mental disorder status is not straightforward. I think that here we should distinguish between two questions: 1. What are the capacities whose impairment causes a harmful condition which can be characterized as a mental disorder? 2. To what degree does a person need to possess these capacities to claim that her condition is not harmful?

The first question represents what might be called the objective aspect of the concept of harm. It might be considered as objective because in Graham's account those are the capacities that all people need to have in order to lead a healthy life. The objective aspect of the concept of harm is important because it delineates mental disorders from problems of living. It is not the case that any harmful condition should be characterized as a mental disorder. As mentioned above, mental disorders are harmful conditions caused by dysfunctions in basic psychological capacities.

In contrast to this, the second question refers to the required degree to which people need to possess these capacities. This aspect of the concept of harm can be construed as subjective because it is likely that there will not be a universally fixed threshold that distinguishes degrees of harm that constitute mental disorders from those that do not. This is because the degrees of harm and their relevance will depend on specific goals and values, which differ greatly from a person to person and their social contexts due to irreducible heterogeneity among people and societies they comprise. Therefore, it is likely that the assessment of the degree to which a person needs to possess the relevant capacity will depend on local contexts and sociocultural norms.

Drawing the distinction between objective and subjective components of harm indicates that not all cases of high-functioning autism would be considered as disorders. From this it follows that individual assessment in relation to a context of living and functioning should be made on a case-by-case basis. For example, in a society which cherishes ideals of extreme individualism and independence, lower abilities of communication and emotional engagement exhibited by high-functioning autistics would not be harmful, or would be harmful to a much lesser extent than they would be in a society where such ideals are not cherished. In a similar vein, due to restrictive and repetitive behavior, which is distinctive for autistics, autism would be less harmful or would not be harmful at all in environments that are structured and demand from people to engage in routine activities. Such environments might involve working on sorting jobs and manufacturing lines. Autistics also might be good at engineering, IT, art and design because they are visually oriented, and they tend to focus on details (Cheriyan et al. 2021; Hayward et al. 2019). Due to outstanding memory, they might perform well at math and library science (Everhart 2020; Cheriyan et al. 2021; Hayward et al. 2019). Autistics also might be very good researchers because they present facts without personal bias due to their tendency to rely on logic and to be unemotional (Cheriyan et al. 2021). Finally, autistics often show strong connections to animals, so they might work as veterinary technicians, dog walkers, zookeepers, livestock handlers, and so on (Prothmann et al. 2009; Reed 2021). In such cases we would have reason to think that autism is much less harmful or is not harmful at all, because in these environments the strengths of the specific autistic individual outweigh other traits that might be associated with maladaptation.

7. *Conclusion*

In this paper, I have reviewed reasons for thinking that autism is a mental disorder. I concluded that severe forms of autism can plausibly be thought of as a mental disorder. I have argued that a general conclusion about the disorder status of high-functioning autism cannot be drawn due to the heterogeneity of autism. I have claimed that in every case of high-functioning autism a specific evaluation of harm should be offered to determine the disorder status of that condition. To elaborate on the procedure by which harm in such cases can be evaluated, I relied on Graham's (2010) list of capacities that are generally needed for leading a healthy life. I argued that some of these capacities could be impaired in the case of high-functioning autism, but whether this is so and to what degree should be determined on a case-by-case basis since the severity and harm of these impairments are likely to be context-dependent.*

References

American Psychiatric Association. 1994. *Diagnostic and Statistical Manual of Mental Disorders: DSM-4.* Washington: American Psychiatric Association.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5.* Arlington: American Psychiatric Association.

* I wish to thank Luca Malatesti and Marko Jurjako for reading previous versions of this article and providing very valuable comments. I am also very grateful to Dora Kunović, who has proofread the article. My research is funded by the Croatian Science Foundation (Project RAD, Grant IP-2018-01-3518 and doctoral scheme DOK-02-2021).

Amoretti, C. M. and Lalumera, E. 2019. "Harm Should Not Be a Necessary Criterion for Mental Disorder: Some Reflections on the DSM-5 Definition of Mental Disorder." *Theoretical Medicine and Bioethics* 40 (4): 321–37. https://doi.org/10.1007/s11017-019-09499-4.

Armstrong, T. 2015. "The Myth of the Normal Brain: Embracing Neurodiversity." *AMA Journal of Ethics* 17 (4): 348–52. https://doi.org/10.1001/journalofethics.2015.17.4.msoc1-1504.

Attwood, T. 1998. *Asperger's Syndrome: A Guide for Parents and Professionals*. London: Jessica Kingsley.

Baron-Cohen, Simon. 1995. *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge: MIT Press.

——— 2009. "Autism: The Empathizing – Systemizing (E-S) Theory. " *The Year in Cognitive Neuroscience:* Ann. N.Y. Acad. Sci. 1156: 68–80.

———. 2017. "Editorial Perspective: Neurodiversity – a Revolutionary Concept for Autism and Psychiatry." *Journal of Child Psychology and Psychiatry* 58 (6): 744–747. https://doi.org/10.1111/jcpp.12703.

Bermúdez, L. J. 2005. *Philosophy of Psychology: A Contemporary Introduction*. London: Routledge

Bingham, R. and Banner, N. 2014. "The Definition of Mental Disorder: Evolving but Dysfunctional?" *Journal of Medical Ethics* 40 (8): 537–542.

Biturajac, M. and Jurjako, M. 2022. "Reconsidering harm in psychiatric manuals within an explicationist framework" *Medicine, Health Care, and Philosophy*. https://dx.doi.org/10.1007/s11019-021-10064-x

Blume, H. 1998. "Neurodiversity". The Atlantic. 30 September 1998.

https://www.theatlantic.com/magazine/archive/1998/09/neurodiversity/305909/.

Bolton, Derek. 2013.. "What is mental illness?" In K. W. M. Fulford, M. Davies, R. G. T. Gipps, G. Graham, J. Z. Sadler, G. Stanghellini and T. Thornton (eds.). *The Oxford Handbook of Philosophy and Psychiatry*. Oxford: Oxford University Press, 434-50

Burgoine, Eyrena, and Wing, L 1983. "Identical Triplets with Asperger's Syndrome." *British Journal of Psychiatry* 143: 261–265.

Cartwright, A. S. 1851. "Report on the Disease and Physical Peculiarities of the Negro Race." *The New Orleans Medical and Surgical Journal* 89–92.

Chapman, R. 2016. "Autism Isn't Just a Medical Diagnosis — It's a Political Identity." Medium. 2016. https://medium.com/the-establishment/autism-isnt-just-a-medical-diagnosis-it-s-a-political-identity-178137688bd5.

———. 2019. "Neurodiversity Theory and Its Discontents: Autism, Schizophrenia, and the Social Model of Disability." In S. Tekin and R. Bluhm (eds.). *The Bloomsbury Companion to Philosophy of Psychiatry*. London: Bloomsbury Academic, 371–390.

Cheriyan, C., Shevchuk-Hill, S., Riccio, A., Vincent, J., Kapp, S. K., Cage, E., Dwyer, P., Kofner, B., Attwood, H., and Gillespie-Lynch, K. 2021. "Exploring the Career Motivations, Strengths, and Challenges of Autistic and Non-autistic University Students: Insights From a Participatory Study". *Frontiers in Psychology* 12: 719–827. https://doi.org/10.3389/fpsyg.2021.719827

Cooper, V. R. 2007. *Psychiatry and Philosophy of Science*. Stockfield: Acumen.

———. 2021. "On Harm." In L. Faucher et D. Forest (eds.). *Defining mental disorders: Jerome Wakefield and his critics*. Cambridge: MIT Press: 537–551.

Cushing, S. 2018. "Has Autism Changed?" In M. dos Santos and J.-F. Pelletier (eds.). *The Social Constructions and Experiences of Madness*. Leiden: Brill: 75–94.

Den Houting, J. 2019. "Neurodiversity: An insider's perspective". *Autism* 23 (2): 271–273

Doherty, J. M. Campbell, M. N., Tsuji, H. and Phillips, A. W. 2010. "The Ebbinghaus Illusion Deceives Adults but Not Young Children". *Developmental Science* 13 (5): 714–721. https://doi.org/10.1111/j.1467-7687.2009.00931.x.

Dominus, S. 2019. "Open Office". *The New York Times Magazine*. Retrieved from https://www.nytimes.com/interactive/2019/02/21/magazine/autism-office- design.html.

Everhart, N., and Anderson, A. M. 2020. "Research Participation and Employment of Persons with Autism Spectrum in Library and Information Science: A Review of the Literature". *Library Leadership & Management* 34 (3). https://doi.org/10.5860/llm.v34i3.7376

Feather, K. A. 2016. "Low functioning to high-functioning autism: A prescriptive model for counselors working with children across the spectrum." *Ideas and research you can use: VISTAS 2016*.

Fletcher-Watson, S. and Happé, F. 2019. *Autism: A New Introduction to Psychological Theory and Current Debate*. https://doi.org/10.4324/9781315101699.

Frith, U. 1989. *Autism: Explaining the Enigma*. Malden: Blackwell.

Garson, J. 2021. "The Developmental Plasticity Challenge to Wakefield's View". In L. Faucher and D. Forest (eds.). *Defining Mental Disorder: Jerome Wakefield and his Critics*. Cambridge: MIT Press: 335–352.

Graby, S. 2015. "Neurodiversity: Bridging the Gap between the Disabled People's Movement and the Mental Health System Survivors' Movement?" In H. Spandler, J. Anderson and B. Sapey (eds.). *Madness, Distress and the Politics of Disablement*. Bristol: Policy Press, 231–43. https://doi.org/10.2307/j.ctt1t898sg.

Graham, G. 2010. *The Disordered Mind: An Introduction to Philosophy of Mind and Mental Illness*. New York: Routledge.

Happé, F. 1999. "Autism: Cognitive Deficit or Cognitive Style?" *Trends in Cognitive Sciences* 3 (6): 216–222. https://doi.org/10.1016/s1364-6613(99)01318-2.

———. 2018. "Why Are Savant Skills and Special Talents Associated with Autism?" *World Psychiatry* 17 (3): 280–281. https://doi.org/10.1002/wps.20552.

Hayward, S. M., McVilly, K. R., and Stokes, M. A. 2019. "Autism and employment: What works." *Research in Autism Spectrum Disorders* 60: 48–58. https://doi.org/10.1016/j.rasd.2019.01.006

Hughes, A. J 2021. "Does the heterogeneity of autism undermine the neurodiversity paradigm?" *Bioethics* 35: 47–60. https://doi.org/10.1111/bioe.12780.

Jaarsma, P. and Welin, S. 2012. “Autism as a Natural Human Variation: Reflections on the Claims of the Neurodiversity Movement.” *Health Care Analysis* 20 (1): 20–30. https://doi.org/10.1007/s10728-011-0169-9.

Jurjako, M. 2019. “Is Psychopathy a Harmful Dysfunction?”*Biology & Philosophy* 34 (5). https://doi.org/10.1007/s10539-018-9668-5

Kanner, L. 1943. “Autistic Disturbances of Affective Contact.” *Nervous Child* 2: 217–250.

Kingma, E. 2013. “Naturalist Accounts of Mental Disorder.” In K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G. Stanghellini and T. Thornton (eds.). *The Oxford Handbook of Philosophy and Psychiatry*. Oxford: Oxford University Press, 363–384.

Lancellotta, E. and Bortolotti, L. 2020. “Delusions in the Two-Factor Theory: Pathological or Adaptive?”*European Journal of Analytic Philosophy* 16 (2): 37–57. https://doi.org/10.31820/ejap.16.2.2

Legault, M., Bourdon, J. N., and Poirier, P. 2019. “Neurocognitive variety in neurotypical environments: The source of “deficit” in autism.” *Journal of Behavioral and Brain Science* 9 (6): 246.

Legault, M., Bourdon, J. N., and Poirier, P. 2021. “From neurodiversity to neurodivergence: the role of epistemic and cognitive marginalization.” *Synthese* 199 (5): 12843–12868.

Lim, C. 2017. “Reviewing Resistances to Reconceptualising Disability.” *Proceedings of the Aristotelian Society* 117 (3): 321–331.

Malatesti, L., Jurjako, M. and Meynen, G. 2020. The Insanity Defence Without Mental Illness? Some Considerations.” *International Journal Of Law And Psychiatry* 71: 101571. doi:10.1016/j.ijlp.2020.101571.

McKay, T. R. and. Dennett, C. D. 2009.“The Evolution of Misbelief.” *Behavioral and Brain Sciences* 32 (6): 493–510. https://doi.org/10.1017/S0140525X09990975

McLaughlin, C. E. and Sutton, J. 2018. “Autistic Man Who Went Overboard on Carnival Cruise Was Traveling with Special Needs Group.” CNN. 2018. https://www.cnn.com/2018/12/20/us/autistic-man-overboard-carnival-cruise/index.html.

McNally, R. J. 2001. “On Wakefield’s Harmful Dysfunction Analysis of Mental Disorder.” *Behaviour Research and Therapy* 39 (3): 309–314.

Meilleur, S. A., Jelenic, P. and Mottron, L. 2014. “Prevalence of Clinically and Empirically Defined Talents and Strengths in Autism.” *Journal of Autism and Developmental Disorders* 45 (5): 1354–1367. https://doi.org/10.1007/s10803-014-2296-2.

Mercier, Céline, Mottron, L. and Belleville, S. 2000. “A Psychosocial Study on Restricted Interests in High-Functioning Persons with Pervasive Developmental Disorders.” *Autism* 4 (4): 406–425.

Meyerding, J. 2014. “Thoughts on Finding Myself Differently Brained.” *Autonomy, the Critical Journal of Interdisciplinary Autism Studies* 1 (3).

Miller, L. B., Cummings, J., Mishkin, F., Boone, K., Prince, F., Ponton, M. and Cotman, C. 1998. “Emergence of Artistic Talent in Frontotemporal Dementia.” *Neurology* 51 (4): 978–982. https://doi.org/10.1212/WNL.51.4.978.

Murphy, D. 2006. *Psychiatry in the Scientific Image*. Cambridge: The MIT Press.

Murphy-Hollies, K. "When a Hybrid Account of Disorder is not Enough: The Case of Gender Dysphoria." *European Journal of Analytic Philosophy* 17 (2): 6–26. https://doi.org/10.31820/ejap.17.3.5

Nelson, H. R. 2021, "A Critique of the Neurodiversity View." *Journal of Applied Philosophy* 38: 335–347. https://doi.org/10.1111/japp.12470

Nicolaidis, C. 2012. "What can physicians learn from the neurodiversity movement?". A*ma Journal of Ethics* 14 (6): 503–510.

Ortega, F. 2009. "The Cerebral Subject and the Challenge of Neurodiversity." *BioSocieties* 4 (4): 425–45. https://doi.org/10.1017/S1745855209990287.

Prothmann, A., Ettrich, C., and Prothmann, S. 2009. "Preference for, and Responsiveness to, People, Dogs and Objects in Children with Autism." *Anthrozoös* 22 (2): 161–171. https://doi.org/10.2752/175303709X434185

Rawls, J. 1971. *A Theory of Justice*. Cambridge: Harvard University Press.

Reed, D. 2021. "Autism spectrum disorder in veterinary clients: How the practice can help." *Veterinary Nursing Journal*, 36 (1): 30–32. https://doi.org/10.1080/17415349.2020.1840472

Reznek, L. 1987. *The Nature of Disease*. London: Routledge and Kegan Paul.

Robeyns, I. 2016. "Conceptualising Well-Being for Autistic Persons." *Journal of Medical Ethics* 42 (6): 383–90. https://doi.org/10.1136/medethics-2016-103508.

Sarrett, C. J. 2016. "Biocertification and Neurodiversity: The Role and Implications of Self-Diagnosis in Autistic Communities." *Neuroethics* 9 (1): 23–36. https://doi.org/10.1007/s12152-016-9247-x.

Sinclair, J. 1993. "Don"t Mourn for Us." *Our Voice* 1 (3). http://www.autreat.com/dont_mourn.html.

Spitzer, L. R. and Endicot, J. 1978. "Medical and Mental Disorder: Proposed Definition and Criteria." In Robert L. Spitzer and Donald F. Klein (eds.). *Critical Issues in Psychiatric Diagnosis*. New York: Raven Press, 15–24.

Spitzer, L. R. and Wilson, T. P. 1975. "Nosology and the Official Psychiatric Nomenclature." *Comprehensive Textbook of Psychiatry* 2.

Stegenga, J. 2015. "Effectiveness of medical interventions." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 54: 34–44.

Stegenga, J. 2021. "Medicalization of sexual desire." *European Journal of Analytic Philosophy* 17 (2): 5–32. https://doi.org/10.31820/ejap.17.3.4

Straus, N. J . 2013. "Autism as Culture." *The Disability Studies Reader* 4: 460–484.

Szasz, T. 1974. *The Myth of Mental Illness*. New York: Harper and Collins

Šustar, P. and Brzović, Z. 2014. "The Function Debate: Between 'Cheap Tricks' and Evolutionary Neutrality." *Synthese* 191 (12): 2653–2671. https://doi.org/10.1007/s11229-014-0407-4.

Treffert, Darold A. 2009. "The Savant Syndrome: An Extraordinary Condition. A Synopsis: Past, Present, Future." *Philosophical Transactions of the Royal Society B: Biological Sciences* 364 (1522): 1351–1357. https://doi.org/10.1098/rstb.2008.0326.

Verhöff, B. 2012. "What Is This Thing Called Autism? A Critical Analysis of the Tenacious Search for Autism's Essence." *BioSocieties* 7 (4): 410–32. https://doi.org/10.1057/biosoc.2012.23.

Wakefield, C. J. 1992. "The Concept of Mental Disorder. On the Boundary between Biological Facts and Social Values." *The American Psychologist* 47 (3): 373–388.

———. 2007. "The Concept of Mental Disorder: Diagnostic Implications of the Harmful Dysfunction Analysis." *World Psychiatry* 6 (3): 149–156.

———. 2014. "The Biostatistical Theory Versus the Harmful Dysfunction Analysis, Part 1: Is Part-Dysfunction a Sufficient Condition for Medical Disorder?" *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 39 (6): 648–682. https://doi.org/10.1093/jmp/jhu038.

Wakefield, C. J. and Conrad, A. J. 2019. "Does the harm component of the harmful dysfunction analysis need rethinking?: Reply to Powell and Scarffe." *Journal of Medical Ethics* 45 (9): 594–596. https://doi.org/10.1136/medethics-2019-105578

Wakefield, C. J., Wasserman, D. and. Conrad, A. J. 2020. "Neurodiversity, Autism, and Psychiatric Disability." In A. Cureton and D. T. Wasserman (eds.). *The Oxford Handbook of Philosophy and Disability*. Oxford: Oxford University Press: 501–521 https://doi.org/10.1093/oxfordhb/9780190622879.013.29.

Wasserman, D., Asch, A., Blustein, J. and Putnam, D. 2016. "Disability: Definitions, Models, Experience". In Edward N. Zalta (ed.). *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/sum2016/entries/disability/.

Weiskopf, A. D. 2017. "An Ideal Disorder? Autism as a Psychiatric Kind." *Philosophical Explorations* 20 (2): 175–90. https://doi.org/10.1080/13869795.2017.1312500.

World Health Organization, ed. 2001. *International Classification of Functioning, Disability and Health: ICF*. Geneva: World Health Organization.

Croatian Journal of Philosophy
Vol. XXIII, No. 67, 2023
https://doi.org/10.52685/cjp.23.67.6
Received: October 5, 2021
Accepted: March 16, 2023

Rawls and the Global Original Position

JINGHUA CHEN
The University of Auckland, Auckland, New Zeland

Cosmopolitans including Charles Beitz, David Richards, Brian Barry, Thomas Pogge and Gillian Brock propose the device of an original global position to work out global principles of justice. However, John Rawls does not agree with this kind of proposal. In this paper, I add two key original contributions, which go beyond previous arguments by cosmopolitans and advance the current debates. First, to argue against Rawls's objection to the global original position, I demonstrate the importance of the distinction between accepting a particular substantive principle and accepting the original position procedure. Second, in order to respond to cultural pluralism, I take a unique approach to show that the idea of the person as free and equal is a fundamental part of the global public culture by examining the most fundamental legal documents: the proto-constitutional documents in international law and the constitutions of the major states. I apply Samuel Huntington's classification of civilisations to identify the major civilisations and their core states and show that the idea of the person as free and equal is implicit in the constitutions of most influential countries even these countries are categorised in different civilisations.

Keywords: Peace; Rawls; the law of peoples; the global original position; cultural pluralism.

1. *Introduction*

The "original position" with its "veil of ignorance" is a model of representation that Rawls designs to develop the political principles of domestic justice as the fair clause of social cooperation in *A Theory of Justice* (Rawls 1999a: 11). The justificatory perspective of the original position, the focus on the basic structure and the selection of guiding principles from the original position are three major theoretical features in Rawls's theory of justice. Rawls envisages a fair social cooperation clause as agreed upon by all those involved in social cooperation. And the consent of free and equal citizens must be under the right conditions. We get the idea of the "original position" by combining the fair conditions to be observed in the formulation of the social fair cooperation clause, that is, the principle of justice. Rawls writes, "I have said that the original position is the appropriate initial status quo which ensures that the fundamental agreements reached in it are fair. This fact yields the name 'justice as fairness'" (Rawls 1999a: 15).

A close relationship exists between the original position and Kantian constructivism (Rawls 1999b: 303). Rawls points out that Kantian constructivism specifies a particular conception of persons as rational agents in a construction procedure according to certain reasonable requirements, which determines the first principles of justice through the agreements of these rational persons (Rawls 2001: 516). Kantian constructivism links the conception of the person, the reasonable procedure of construction and the principles of justice. For Rawls, Kantian constructivism is the best way to justify a proper conception of justice we can hope for. Moral objectivity is not independent of the social or human point of view. It can only be constructed (through a procedure) on the acceptable moral facts by free and equal, reasonable and rational moral persons.

In his international theory in *The Law of Peoples*, nevertheless, Rawls criticises and rejects the approach of the global original position. He applies the device of the original position in two stages (involving three uses) rather than a single global original position for the selection of political principles of international society. Those represented in the second stage of the original position are peoples, who are collective entities rather than individual persons. In the first step of the second stage, the liberal peoples agree upon the law of the peoples for the society composed of liberal peoples. In the second step of the second stage, the liberal peoples propose to the decent peoples the selected eight principles of the law of the peoples in the previous step. Rawls claims that decent peoples would accept these eight principles in good faith. Thus, Rawls gives up the idea of the person as free and equal, reasonable and rational individual as the justificatory foundation to work out the international principles, and hence deviates from Kantian constructivism in his international theory.

Rawls's shift in his approach has significant theoretical implications and consequences in international theory. This paper attempts to sort outs Rawls' objections against the global original position and then present corresponding analyses and responses. It also compares the theoretical advantages and disadvantages of the global original position and Rawls's approach in *The Law of Peoples*. Hopefully, this study may shed some light on clarifying justificatory grounds in selecting guiding principles for global peace and justice.

2. *The idea of the global original position in the theories of cosmopolitan justice*

Soon after Rawls published *A Theory of Justice* in the 1970s, some other scholars advocated extending the original position to the global context and envisioned a single global original position to reflect on the principles of global justice. These advocates are called cosmopolitans or theorists of cosmopolitan justice.

Charles Beitz writes, "Thus the parties to the original position cannot be assumed to know that they are members of a particular national society, choosing principles of justice primarily for that society. The veil of ignorance must extend to all matters of national citizenship, and the principles chosen will therefore apply globally" (Beitz 1999: 151). Beitz maintains that, once properly reinterpreted, Rawls's two principles of justice can be applied globally. Not only is the state a social cooperation system, but the entire human society is also a global social cooperation system due to increasing economic and political interdependence. The familiar reasoning in Rawls' domestic theory of justice can be applied in the global case. Therefore, Beitz proposes to envisage a single global original position in which contracting parties represent each individual, instead of the state, on a global scale. They choose the principles of global justice behind the veil of ignorance of individual persons' fundamental interests (Beitz 1999: 143–161).

Thomas Pogge also endorses Beitz's proposal and proclaims that nationality is of no moral significance. "Nationality is just one further deep contingency (like genetic endowment, race, gender, and social class)" (Pogge 1989: 246). He envisions a single global original position to construct a global institutional scheme (Pogge 1989: 246, 247, 256).

In addition to Beitz and Pogge, David Richards (1982), Brian Barry (1973, 1989) and Gillian Brock (2009) also support the use of the global original position. This device's primary feature is that all individuals of humankind are considered free and equal. Their consent for the right reasons should ground the justification of the principles of justice for the global basic structure. The global original position embodies moral universalism, which means each individual of humankind has a global stature as the ultimate unit of moral concern (Pogge 2008: 169). Put another way, the contractarian framework of the global original position is based on the concept of "moral reciprocity," that is, "treating persons one would oneself reasonably liked to be treated" (Richards 1982: 281–282).

3. *Rawls's criticisms of the global original position*

Regarding the theoretical device of the global original position proposed by these writers of cosmopolitan justice, Rawls argues against it in both *The Law of Peoples* in 1993 and 1999. In the former work, he claims that the approach of the global original position makes the foundation of the law of the people too narrow. He also enlists the main reasons why the approach in *The Law of the Peoples* is superior to the global original position. The clear defence of his approach includes the following: First, the theory of domestic justice focuses on the basic structure of society, and so far, everything has progressed well. So when formulating the law of the peoples as the guideline for international relations between peoples, it is reasonable to presume the existence of domestic societies and the principles of justice for their basic structure as a starting point. Second, peoples as sovereign entities now exist in some form worldwide. Third, his approach can consider factors such as peoples' considerations and government's consent (Rawls 1999c: 535–536).

In the 1999 work, at least literally, his principal and almost sole reason, which he gives explicitly, for opposing the global original position seems to be that the global original position may lead to global liberal principles of justice and hence the liberal foreign policy, which is unacceptable. For Rawls, the device of the global original position means that all people will have equal rights to liberties owned by citizens of a constitutional democratic society. According to this interpretation, the foreign policy of liberal peoples, which Rawls hopes to clarify, will be a step-by-step approach to shaping all nonliberal societies and moving them towards liberalism. Rawls rejects this kind of foreign policy because it assumes that only a liberal democratic society is acceptable.

After summarising these two major objections to a global original position, I now turn to detailed analysis and responses to them. And I will start with the latter objection because it is Rawls's last opinion concerning the global original position.

4. *The global original position and liberal foreign policy*

4.1. *Liberal rights and the liberal foreign policy*

Rawls objects to the use of the global original position. His primary rationale for his objection is this:

> To proceed in this way, however, takes us back to where we were in ζ7.2 (where I considered and rejected the argument that nonliberal societies are always properly subject to some form of sanctions), since it amounts to saying that all persons are to have the equal liberal rights of citizens in a constitutional democracy. On this account, the foreign policy of a liberal people--which it is our concern to elaborate--will be to act gradually to shape all not yet liberal societies in a liberal direction, until eventually(in the ideal case) all societies are liberal. But the foreign policy simply assumes that only a liberal democratic society can be acceptable. Without trying to work out a

> reasonable liberal Law of Peoples, we cannot know that nonliberal societies cannot be acceptable. (Rawls 1999c: 82–83)

Evidently, his objection resorts to his opposition to liberal foreign policies. Why does Rawls oppose the foreign policy of liberal people to shape all nonliberal societies in a liberal direction? His main argument is that, out of respect for reasonable pluralism, liberal societies should be tolerant of decent societies. If decent societies are not made equal and *bona fide* members of the Society of Peoples, they do not receive due respect. We must clarify whether Rawls is against all people possessing equal rights to liberties or only against liberal foreign policies.

Rawls makes it clear that if a liberal constitutional democracy is indeed better than other forms of society, which he believes is true, then liberal peoples should also believe and assume that once liberal peoples treat decent peoples with due respect, decent societies will gradually recognise the advantages of the free system and take initial actions to make their system freer. He hopes that dissenters in decent peoples will promote the liberal change of decent people (Rawls 1999c: 61). From this point of view, Rawls does not generally oppose that citizens in all societies have the right to equality and freedom. He is only opposed to the liberal diplomatic policy, that is, the adoption of step-by-step measures to shape all nonliberal societies according to the model of liberalism.

4.2. *The distinction between the global original position as a theoretical device and substantive political principles*

Rawls believes that everyone will have equal rights to liberties enjoyed by each citizen in the constitutional democratic society means the need for liberal peoples to pursue a liberal foreign policy. This direct link seems problematic. There are other approaches to promote individuals' rights and liberties in the world. The primary example is the human rights approach adopted by the United Nations. The iconic event is the signing and ratification of the Universal Declaration of Human Rights and other human rights covenants. Another example is the European Union's approach, stipulating conditions for accession to encourage countries interested in joining the EU to become more liberal.

Rawls might respond to the arguments above like this: the international human rights approach and the EU approach are also initiated and executed by democratic countries and hence still part of their foreign policies. But it is proper to insist that these approaches are not unacceptable because both the international human rights approach and the EU model are based on the consent of the participating countries. In Rawls's international theory, only liberal peoples are societies with a genuinely normative feature. Decent peoples are qualified to be tolerated only because they satisfy some of the liberal conditions, such as human rights protection and political consultation. Rawls still hopes that decent peoples will eventually implement liberal reforms moved by domestic dissenters. His opposition to the liberal foreign policy is essentially against the compulsory liberalisation of decent peoples by foreign regimes. Nonetheless, decent peoples can achieve liberalisation voluntarily under the international and EU human rights approaches. Thus, it is not tenable for Rawls to link liberal rights and liberties directly to the liberal foreign policy of democratic countries.

More importantly, Rawls's criticism of the global original position does not distinguish between the global original position as a theoretical device and the principles of justice derived from the original position and the resulting foreign policy. He believes that starting from the global original position would necessarily lead us to conclude with the choice of liberal foreign policy. But this connection is untenable. As Rawls states in *A Theory of Justice*, the original position and the principle of justice are two separate parts of the contractual theory of justice. A person can agree to the original position without agreeing to the specific principles of justice derived therefrom, and vice versa. Rawls writes:

> It is, therefore, worth noting from the outset that justice as fairness, like other contract views, consists of two parts: (1) an interpretation of the initial situation and of the problem of choice posed there, and (2) a set of principles which, it is argued, would be agreed to. One may accept the first part of the theory (or some variant thereof), but not the other, and conversely. The concept of the initial contractual situation may seem reasonable although the particular principles proposed are rejected. (Rawls 1999a: 14)

Therefore, even if the opposition to a particular substantive principle and its practical implication is sound, it cannot necessarily constitute an effective rebuttal to the global original position as a model of representation.

It is important to note that if the veil of ignorance of the global original position or other relevant supporting conditions is modified, the principles of justice obtained in this global original position may vary. What principles of justice will be derived from the device of the original position depends on the setting of the veil of ignorance and the interpretation of the relevant conditions. Therefore, the device of the global original position per se does not necessarily lead to a particular principle of justice. And the objection of the use of the global original position cannot be justified on the ground of unacceptable liberal foreign policy.

5. *Cultural pluralism in the law of peoples*

5.1. *Challenge from cultural pluralism in working out global political principles*

In the 1993 paper The Law of Peoples, Rawls's reason for opposing the global original position is that it makes the foundation of the law of peoples too narrow. Rawls suggests the trouble with the global original position, which is all-inclusive, is that it has many problems with the use of the concept of freedom; because in the global case, the global original position is meant to treat all people, regardless of their society and culture, as free and equal, reasonable and rational individuals in order to conform to the concept of liberalism. This makes the foundation of the law of peoples too narrow. By this, he means that the global original position envisions people as free and equal, hence does not tolerate the perspective of a nonliberal society. Liberal peoples should not impose our own culture and values on decent peoples. The value of nonliberal society should be respected and tolerated equally. In decent societies, especially those organised by a comprehensive religious, moral or philosophical doctrine, people do not regard each other as free and equal. Therefore, presuming that all people are free and equal is unacceptable (Rawls 1999a: 549–550). Although two objections are closely interconnected, they are not entirely the same. In the previous objection, Rawls objects to the global original position because of its consequence: the unacceptable liberal foreign policy. By contrast, the essence of the latter objection is that the concept of person per se embodied in the global original position is troublesome. This contention is based on cultural pluralism.

5.2. *Global public political culture*

As Pogge maintains, criticism from cultural pluralism is the most serious objection to the globalisation of Rawls's principles of justice. He writes, "We must not impose our values upon the rest of the world, must not pursue a program of institutional reform that envisions the gradual supplanting of all other cultures by a globalised version of our own culture and values. This is, I think, the most serious objection to globalising Rawls and the one that seems to have influenced Rawls himself" (Pogge 1989: 267).

Although cultural pluralism ought to be respected, a universal concept of person is indispensable even in Rawls's *The Law of Peoples* and in the theories of some writers who oppose the globalisation of Rawls's principles of domestic justice.

In order to set up the criterion of decency of societies and hence determine corresponding foreign policies of liberal democratic countries, Rawls proposes a thin list of human rights in *the Law of Peoples.* The protection of basic human rights for every human being is still regarded as a universal starting point. That is to say, in the aspect of the protection of basic human rights, Rawls regards everyone in different societies as equal. Also, although Thomas Nagel objects to globalising the principle of distributive justice, he advocates that we have minimal concerns about human compatriots who have long been suffering from hunger or severe malnutrition, and died from preventable diseases (Nagel 2005: 118). This concern does not need to be predicated on the existence of a special relationship, but only on the humanity we share.

The claim for basic human rights is equal for every human being.

More importantly, Rawls approves the important role of liberal dissenters within decent peoples and hopes that they will promote decent peoples to freedom. Therefore, the concept of the person as free and equal in the global original position is consistent with Rawls's hope that the decent societies will reform in the direction of liberalism.

In addition, although Michael Blake agrees with Rawls's two-tiered principles of justice, that is, the international society is different from the domestic society, his argumentation can be said to be consistent with the core value of the global original position. He asserts that every human being has autonomy; the selection of different principles of justice depends on the different relationships that exist between people. This way of selecting the principle of justice can fulfil the requirement of impartiality to everyone (Blake 2001: 265-273, 281-285). Blake's argument exemplifies that building the principles of global justice by treating everyone as free and equal can express more profound respect for individuals in other societies.

Cultural pluralism does not mean moral relativism. Against the background of pluralism, the ideas of basic human rights, freedom, autonomy or humanity may still reasonably serve as the fundamental principles in reaching a global overlapping consensus.

Nevertheless, it is disputable on what these fundamental ideas should be. Leif Wenar and Amy Eckert, in order to defend Rawls' choice of the international original position, argue that, as in Rawls's *Political Liberalism* the principle of domestic justice is grounded in the fundamental ideas of the public political culture in the constitutional democratic society, Rawls follows the same idea in the issue of global justice, that is, relying on the fundamental ideas in the global public political culture in order to formulate the overlapping consensus as the global guiding principle. Since the Westphalia Peace Treaty more than three hundred years ago, we can see from the practice of international treaties, customs and international organisations that peoples (or states) rather than individuals are the main political actors. Although many countries have signed various international human rights declarations and conventions since World War II, their implementation still depends on the states (Eckert 2006: 851). So when formulating the principles of international justice, it is more appropriate to represent the peoples or the states in the original position. Moreover, Leif Wenar holds that Rawls's approach in international justice is superior to the global original position. He questions the global original position's capacity to develop the necessary principles of international relations like "nations should keep their treaties" (Wenar 2002: 72).

Nevertheless, even assuming that Eckert and Wenar's interpretation of Rawls is sound, that is, the formulation of the principle of global justice requires finding some fundamental ideas that can ground a global consensus in the global public political culture, it is not unreasonable for us to believe that after the Enlightenment, especially after World War II, freedom and equality of individual persons have gradually become universal values.

Rawls builds up his principles of justice based on the fundamental ideas implicit in the public political culture of constitutional democracy. Public political culture is reflected in the Constitution, constitutional documents and judiciary reviews. Following a similar approach, we may identify the fundamental ideas implicit in the global public culture by examining the "proto-constitution" (Habermas 2006: 133) in international society and the constitutions in the major states.

The UN Charter and the International Bill of Human Rights are widely regarded as proto-constitutional documents in the international dimension. The idea of the person as free and equal is explicit in these essential documents. The UN charter expresses the ends of establishing the UN in the opening. The second end is "to reaffirm faith in fundamental human rights, in the dignity and worth of the human person, in the equal rights of men and women and of nations large and small ···" Freedom and equality are two fundamental values manifest in the UN charter. This becomes more explicit in the International Bill of Human Rights, which consists of the Universal Declaration of Human Rights (1948), the International Covenant on Civil and Political Rights (ICCPR, 1966) and the International Covenant on Economic, Social and Cultural Rights (ICESCR, 1966). The idea of the person as free and equal is widely endorsed by these covenant parties.

More importantly, this judgment can be confirmed by domestic constitutions, which are the most fundamental legal documents making up their public political culture. In the following discussions, I will demonstrate that the idea of the person as free and equal is implicit in the constitutions of most influential countries even these countries are categorised in different civilisations.

What are these major civilisations? And who are these major countries? To identify them, I will exploit one of the most prominent IR theorists in the Post-Cold War era, Samuel Huntington's classification of civilisations. According to Huntington, there are eight or nine major civilisations in the world. And these civilisations have their own core states and corresponding concentric circles. After sorting out Huntington's text in *The Clash of Civilizations and the Remaking of World Order, here is the list of these civilisations and core states* (Huntington 1996: 45–48, 155–179).

Civilisations	Core States
Western or Christian	USA (and Europe)
Japanese	Japan
Hindu	India
Latin civilisation	Mexico, Brazil, Argentina
African	Probably South Africa
Sinic or Confucian	China
Orthodox	Russia
Islamic	Indonesia, Egypt, Iran, Pakistan, Saudi Arabia (I would add Turkey, Malaysia)
Buddhist	I would enlist Vietnam, Thailand, Sri Lanka and Myanmar.

The USA, Europe, Japan, India, Mexico, Brazil, Argentina, and South Africa have a democratic political system even though they belong to Western, Japanese, Hindu, Latin, and African civilisations. I choose to investigate those countries whose constitutional nature is not so obvious: Egypt, Iran and Saudi Arabia in the Islamic civilisation, Vietnam, Thailand and Myanmar in the Buddhist civilisation, in addition to China and Russia.

The major findings are put in the following table:

	China	Russia	Egypt	Iran	Saudi Arabia	Vietnam	Thailand	Myanmar
General stipulation on the equal right and freedom	Article 33	Article 17, 19, 21, 22, 45	Article 51, 92, 93	Article 19	Article 26	Article 50	Article 4, 25	Article 348, 353
equality before the law	33	19	53	20		52	27	347
freedom of the person	37	20, 21	54, 55, 59, 60	22, 38, 39		71	28	353
freedom of speech, the press, assembly and association	35	29, 30, 31	65, 70, 71, 73, 74, 75, 76,77	24, 26, 27		69	34, 35, 42, 44, 45	354
freedom of conscience and religion	36	28	64	23		70	31	34, 362, 363
electoral right (the right to participate in managing state affairs)	34	32	87	58, 59		53, 54	2, 50	369

All these countries surveyed except Saudi Arabia stipulate the basic principle of equal freedom and the basic rights and liberty in their constitutions. Even Saudi Arabia's basic law refers to the protection of human rights.

To conclude, the idea of the person as free and equal is implicit in the global public political culture. This is reflected not only in various international and regional human rights declarations and human rights conventions but also in most influential countries' constitutions. In this historical situation, it is not unreasonable to formulate a global overlapping consensus starting from the idea of treating each individual as free and equal, implicit in the global public political culture.

6. *Comparing the global original position and Rawls's approach in international justice*

In the previous sections, I have laid out defensive arguments to address Rawls' objections to show that these oppositions cannot effectively refute a global original position as a permissible model of representation in developing principles of global peace and justice. Now I turn to the offensive arguments. In this section, I compare the global original position and Rawls's approach in international justice and try to demonstrate the superiority of the former. First of all, I discuss the ultimate aim and priority of the theme in *The Law of Peoples* to show that global peace and stability is the dominant theme in *the Law of Peoples*. Given this specific theme or theoretical goal, I will expose the theoretical dilemmas of Rawls's approach and the theoretical advantages of the global original position. I will argue that the openness of the global original position is a significant advantage, for the device of global original position can allow us to consider more alternatives than the international original position does.

6.1. *The ultimate aim in* The Law of Peoples

The ultimate aim of Rawls's theoretical construction is to indicate the direction in the global order to eliminate the great evils of human history (Rawls 1999c: 6–7) and guarantee that "peace and justice would be achieved between liberal and decent peoples both at home and abroad" (Rawls 1999c: 6). Rawls pins the hope for a realistic Utopia in democratic societies (liberal peoples). He claims: "Our hope for the future of our society rests on the belief that the nature of the social world allows reasonably just constitutional democratic societies existing as members of the Society of Peoples" (Rawls 1999c: 7). It is useful to bear this aim in mind while comparing Rawls's approach and a global original position.

6.2. *Rawls's first theoretical dilemma: Unreasonable assumption of altruistic motivation of well-ordered societies and the establishment of a realistic utopia*

In *The Law of Peoples*, the original position is used three times to formulate the eight principles of the Law of Peoples. These eight principles have been devised by representatives of liberal peoples, although they must consider that the principles are reasonably acceptable to decent peoples. From the discussions above, it can be seen that the main rationale why Rawls starts with liberal peoples is that he pins the hope for a peaceful and stable world upon them. Based on this conviction, the Law of Peoples revolves around how liberal peoples should treat decent peoples, outlaw states, and burdened societies. On the other hand, the ultimate concern of the Law of Peoples is to establish a realistic Utopia to eliminate the great evils in human history. If this goal's realisation relies on liberal peoples, it presumes an unreasonable motivation of liberal peoples. Three major foreign policies as the means of realising the Society of Peoples include liberal toleration of decent peoples, intervention in outlaw states on the ground of gross violations of human rights, and duty of assistance towards burdened societies. The last two foreign policies assume altruism of liberal democracies and decent peoples.

But Rawls claims in *A Theory of Justice*, "At the basis of the theory, one tries to assume as little as possible" (Rawls 1999a: 110). By this, he means it is too strong to assume that the motive of the representatives of individuals is altruistic. This is also true for liberal peoples and decent peoples. In the case of liberal peoples, they are political societies established to benefit the citizens. They are hence self-interested. According to Rawls, there are three primary characteristics of liberal peoples: first, their fundamental interests are served by a reasonably just constitutional democratic government; second, citizens are united through what Mill called "common sympathies"; third, peoples have a certain moral character. The first feature is institutional, the second is cultural, and the third is moral and requires a firm attachment to a political (moral) concept of right and justice (Rawls 1999c: 24–25). Rawls writes, "As reasonable citizens in domestic society offer to cooperate on fair terms with other citizens, so (reasonable) liberal (or decent) peoples offer fair terms of cooperation to other peoples. A people will honour these terms when assured that other peoples will do so as well" (Rawls 1999c: 25). He also states that democratic societies are self-satisfied and have no reason to violate other countries. But even in such an idealised definition, we do not see that liberal peoples have the altruistic motive to intervene or assist other states and hence eliminate the great evils of humanity in the long run.

Moreover, Rawls acknowledges that the United States, a constitutional democracy, has repeatedly unjustly overthrown other governments, even though these countries have established some aspects of democracy. He writes, "Hence, given the great shortcomings of actual, allegedly constitutional democratic regimes, it is no surprise that they should often intervene in weaker countries, including those exhibiting some aspects of a democracy, or even that they should engage in war for expansionist reasons. As for the first situation, the United States overturned the democracies of Allende in Chile, Arbenz in Guatemala, Mossadegh in Iran, and, some would add, the Sandanistas in Nicaragua" (Rawls 1999c: 53). This historical evidence in international relations raise doubt on the reliability of the goodwill of constitutional democracies. Therefore, if liberal peoples are assumed to own the motivation to establish a realistic Utopia to eliminate the world's great evils, it is contrary to empirical evidence. And if a liberal people does not have such an altruistic motive, how can it eliminate the world's great evils with its foreign policy? Philip Pettit expresses similar scepticism. He writes, "If there is a weakness in Rawls's schema it shows up, ironically, with the principles on which radical cosmopolitans are likely to agree rather than disagree: namely, that well-ordered peoples should help those who live under oppressive and burdened regimes. If those in the second original position represent only well-ordered societies and not individuals across all societies, then it is unclear why they would have a rational motive for endorsing such altruism" (Pettit 2006: 54).

The asymmetry between the motive of well-ordered peoples and the purpose of eliminating the great evils in human history is the first dilemma that is difficult to overcome in Rawls's approach in formulating the law of the peoples.

Furthermore, in Rawls's thought experiment, the eight principles of the Law of Peoples are developed by representatives of liberal peoples, whereas decent peoples have no right to propose the principles. In Rawls's procedure, decent peoples are not situated symmetrically with liberal peoples, not to mention burdened societies, benevolent absolutisms and outlaw states. Kok-Chor Tan contends that since decent hierarchical societies are not democratic, they cannot be represented reasonably in the original position (Tan 1998: 286–287). Therefore, the formulation of the Law of Peoples is rather like the legislation for the world by liberal peoples. This kind of unilateral legislation is contrary to the core position of contractualism, which Rawls claims to apply.

6.3. *Rawls's second theoretical dilemma: Starting from sovereign state system and reflection on the global basic structure*

The weakness in Rawls's approach discussed above might also be made up in another way in which all the political societies are situated symmetrically in the original position. Such an international original position seems to be able to avoid the unreasonable assumption of an altruistic motivation, and achieve greater allegiance and stability. However, I will argue that, even though impartial for all the states, this international original position still encounters another serious theoretical dilemma.

This formulation process is difficult to truly reflect on the sovereign state system to submit sovereignty to the interests of humankind, that is, to eliminate the great evils in human history. The following arguments can also be seen as targeted against the theoretical approach in *The Law of Peoples*, for liberal peoples are also a kind of political society with sovereignty.

On the one hand, the state system helps maintain order internally and resist external aggression. On the other hand, there is also preliminary evidence to suggest that the state system also has a major adverse effect in life and property through wars, armed conflicts and other politically organised violence. Kant points out the double-edged feature of state sovereignty. Brown writes, "As will be discussed below, Kant wants to challenge the natural law doctrine supporting state sovereignty while also dismissing arguments advocating the creation of a world state. In this regard, Kant's international theory tries to navigate a middle passage between the idea that states can act as the ultimate protectors of human freedom, while also aware of the fact that states are often the primary violators of this very freedom" (Brown 2009: 89). Andrew Kuper also opines, "The horrors of nationalistic wars, xenophobia, and unnecessary starvation might motivate instead a greater focus on human individuals regardless of their geographical location and-as Pogge argues-on lowering the stakes (and hence incentives to abuse) that attach to each institutional level and domain. If history suggests anything, it is that we should scrupulously interrogate and dismiss assumptions that might be destructively 'trapping us in the buildings and boundaries' of the past or present" (Kuper 2000: 660). Therefore, there is sufficient reason to reflect upon the state system in order to make it yield to the interests of humankind.

The international original position will encounter a paradox in this kind of moral reflection because the moral personality of states makes it self-contradictory to adjust or transform sovereignty. To be specific, if the goal of the representatives of states is to determine principles of justice as the fair clauses of social cooperation between countries, it is self-contradictory to constrain state sovereignty by the execution of the principle of justice, because it means to undermine the moral personality of the states. Put another way, states execute the contract between states, so preserving the state's moral personality is logically necessary. Just like the case of deliberation between individuals, it is self-contradictory for the contracting parties to achieve an agreement in which contracting parties become persons with no or limited capacity for civil conducts. Likewise, the international original position cannot seriously reflect on and adjust state sovereignty system. It is hence not suitable to consider more important proposals for global peace and eliminate great evils in human history.

Pogge writes, "In Rawls's sketch, the mere existence of states system in its current form reduces the agenda of the parties' global session to dealings between governments and motivates the priority of domestic over global principles of justice. His endorsement of this institution can have force, however, only if it has been subjected to moral examination (like other social institutions). Otherwise, Rawls would be begging a crucial question, provided we allow, as reasonably we must at the outset, that justice may fail to require the states system in its present form" (Pogge 1989: 257–258). Although his remark is targeted against Rawls's ideas on international justice in *A Theory of Justice*, it is also a pertinent appraisal on the international original position in *The Law of Peoples*.

Many vital proposals in the history of political theory and current contemporary scholarship, including legal pacifism (Kelsen 1944), cosmopolitan democracy (Held 1995, Archibugi 2008), a subsidiary world republic (Höffe 2007) and constitutionalisation of international law (Habermas 2006: 115–193) challenge the state system. To some extent, these proposals of global order are designed to go beyond the sovereign states system and consider how sovereignty is tamed and prevent evils related closely to the state system.

Unfortunately, due to the constraint of the moral personality of peoples or states, the international original position with representatives of peoples cannot reflect sufficiently upon the state system in order to help eliminate the great evils in human history. The reflection by the sovereign entities presumes the existence of a certain kind of sovereign entities and a particular kind of sovereign state system. But the requirement of global peace and justice needs us to reflect upon the sovereign system per se.

6.4 *Theoretical advantages of the global original position*

Concerning the first dilemma discussed above, the problem of unreasonable motive and the problem of stability can hopefully be avoided in the application of the global original position. The individual contracting parties are contracting with each other for their own benefit with reasonable moral constraints. An altruistic motive need not and should not be presumed. In such a procedure, there is no exorbitant requirement that parties must contribute to the well-being of all. The fundamental interests of the represented can be guaranteed rather through formulating fair clauses of cooperation, based on rationality and reasonableness of all parties, than through altruism of one or some particular parties. Also, in the global original position the represented is every individual, the principles of global justice must be justified to all persons with the same reason, and affect them equally. Hence, the principles of justice developed from this procedure can win allegiance and stability more firmly.

Concerning the second dilemma, the global original position approach has a distinct advantage: openness. It does not presume the justice of the status quo and hence can help us exclude the arbitrary moral factors from the existing global system in formulating a global political principle for the global basic structure. The device of the global original position can be used to consider more alternatives than Rawls's international original position, which regards the states system as a starting point for moral reasoning and makes the range of alternatives of global political principles narrower. The global original position is more suitable to consider the historical and present proposals of the world order because the representatives of individuals in a global original position do not necessarily adhere to any existing sovereign state systems. Legitimacy and justice of these systems ultimately need to be justified by appealing to the fundamental interests of every individual. What features sovereign entities should have cannot be determined by the Rawlsian international position, which presumes particular characteristics of the peoples as collective entities. The determination of proper characteristics of peoples is only hopeful to be worked out successfully from the starting point of moral individualism and universalism embodied in the device of the global original position.

To be more specific, starting our arguments from the global original position helps us consider more alternatives concerning the global basic structure, such as the proposals of world government and realism. In contrast, beginning from the original position populated by "peoples", which are sovereign entities with limited sovereignty, would make both of them unqualified as alternatives. From the perspective of Rawls's international original position, realism will be excluded from the beginning because realism presumes absolute sovereignty, which contradicts the characteristic of limited sovereignty of Rawls's "peoples". And the proposal of world government will also be neglected because it means the disappearance of other sovereign entities, such as Rawls's "peoples". This also explains why Rawls claims there are no other alternatives to compete with his eight principles of the law of peoples. The global original position helps us consider the most significant alternatives and can better serve as a legitimate justificatory foundation for comparing and selecting global political principles.

Rawls writes, "As mentioned earlier, the law of peoples might have been worked out by starting with an all-inclusive original position with representatives of all the individual persons of the world. In this case the question of whether there are to be separate societies, and of the relations between them, will be settled by the parties behind a veil of ignorance. Offhand it is not clear why proceeding this way should lead to different results than, as I have done, proceeding from separate societies outward. All things considered one might reach the same law of peoples in either case" (Rawls 1999b: 549). Even this is possible; it is still better that the proper political principles for the global basic structure to be worked out from justifications to individual persons rather than peoples or states. As Kuper argues, "To the extent that the moral claims of states have any normative force in liberalism, it is derivativeit must be justified. In political liberalism, we do not close off the possibility that parties representing free and equal persons in a global original position would decide in favour of thin states or even in favour of an inferior position for a woman within a particular state (although I doubt they would); rather, we say that thin states, and her occupying this position, must be justified" (Kuper 2000: 652).

The philosophical distinction between the global original position and Rawls's international original position can be illuminated by the distinction between cosmopolitan liberalism and social liberalism made by Beitz in his paper "Social and Cosmopolitan Liberalism." Beitz claims that social liberalism advocates that international justice is fundamentally a matter of fairness to societies (or peoples). In contrast, cosmopolitan liberalism insists that this is a matter of fairness to individuals (Beitz 1999: 515). In other words, social liberalism gives an independent ethical status to domestic-level societies, while cosmopolitan liberalism regards individual well-being as fundamental, and the value of society is derived only based on personal interests (Beitz 1999: 520).

Beitz proposes that if social liberalism considers the independent moral value of the state (or society) only because it is the most effective political mechanism that can guarantee human rights, then there is no difference between the two doctrines (Beitz 1999: 529). If individual interests can be merged into the interests of the state, then it seems that the same results can be obtained either with the international original position or the global original position. But as Kuper argues, personal interests cannot be fully incorporated into national interests. He cites immigration between underdeveloped (U) and developed (D) countries as an example. "It might be rational for D to restrict immigration because it would result in a loss of capacity to secure the rights and well-being of its citizens; and it might be rational for U to restrict emigration for similar reasons" (Kuper 2000: 646). He continues, "It may be the case that allowing some more movement of people between the two would result in a gain for those who are worst off or even in a more extensive scheme of basic liberties for all. This is not, however, a consideration that could count for parties representing U and D (sets of citizens) but only for parties representing all the persons in U and D as individual persons" (Kuper 2000: 646). And it is also worth noting that allowing immigration has important interests for the immigrants themselves and their families. This consideration is also difficult to count for representatives of states or peoples. It can be seen that personal interests cannot be fully integrated into national interests. Furthermore, the interests of non-democratic societies (including the decent societies constructed by Rawls) can hardly be said to be able to merge personal interests, especially those interests represented by dissenters. Therefore, the international original position cannot incorporate the global original position.

7. *Concluding remarks*

As we know, the arguments of the original position concerning the selection of the best political principle for the basic structure of society include not only the requirement of morality but also the realistic considerations, which means to evaluate the feasibility and efficiency of the candidate principles through taking account of information of relevant facts, empirical theories and historical experience. The full justification of the proper conception is related to moral constraints and all relevant general facts and theories. The idea of constructivism needs to identify which facts are relevant from the appropriate point of view and to determine their weight as reasons" (Rawls 2000: 246). This makes the justification susceptible to a broad range of arguments, including moral and realistic considerations, and permanently open to criticisms and revisions. The correct moral judgment must be made by agents who are not only reasonable but also fully informed (Rawls 2000: 244).

Just as the original position and Kant's Categorical Imperative procedure, the global original position can also be regarded as attempting to extend the limits of practical possibility realistically towards a moral ideal. The major cosmopolitans, such as Kant, Hans Kelsen and Jürgen Habermas, are well aware of the limits of reality and try to figure out a realistic proposal after considering the particular circumstances of the contemporary situation and the complex historical momentum. They all advocate gradualism rather than revolutions, which can substantially alleviate the worry of "too utopian".

The cosmopolitan project is not necessarily a task that must be accomplished in the near future. Yet, this model provides an appealing, logical and self-sufficient ideal to guide humankind's long-termed endeavours. It may be postponed, and it may even not be realised completely due to the "crooked" half of human nature and human society. Still, it must be recognised and pursued as an ideal that stimulates the arduous efforts of you and me, here and now.

References

Archibugi, D. 2008. *The Global Commonwealth of Citizens: Toward Cosmopolitan Democracy*. Princeton and Oxford: Princeton University Press.

Barry, B. 1973. "The Liberal Theory of Justice: A Critical Examination of the Principal Doctrines in a Theory of Justice by John Rawls." *Philosophical Review*, 84 (4), 598–603.

Barry, B. 1989. *A Treatise on Social Justice Vol. 1*: *Theories of Justice*. Berkeley: University of California Press.

Beitz, C. 1979. *Political Theory and International Relations*. Princeton: Princeton University Press.

Beitz, C. 1999. "Social and Cosmopolitan Liberalism." *International Affairs* 75 (3): 515–529.

Blake, M. 2001. "Distributive Justice, State Coercion, and Autonomy." *Philosophy and Public Affairs* 30 (3): 257–296.

Brock, G. 2009. *Global Justice: a Cosmopolitan Account.* New York: Oxford University Press.

Brown, G. W. 2009. *Grounding cosmopolitanism: From Kant to the idea of a cosmopolitan constitution.* Edinburgh: Edinburgh University Press.

Eckert, A. 2006. "Peoples and Persons: Moral Standing, Power, and the Equality of States International." *Studies Quarterly* 50 (4): 841–859.

Habermas, J. 2006. *The Divided West.* Cambridge: Polity Press.

Held, D. 1995. *Democracy and the Global Order: from the Modern State to Cosmopolitan Governance.* Cambridge: Polity Press.

Höffe, O. 2007. *Democracy in an Age of Globalization.* Dordrecht: Springer.

Huntington, S. P. 1996. *The Clash of Civilizations and the Remaking of World Order.* New York: Simon & Schuster.

Kelsen, H. 1944. *Peace Through Law.* Chapel Hill: University of North Carolina Press.

Kuper, A. 2000. "Rawlsian Global Justice: Beyond the Law of Peoples to a Cosmopolitan Law of Persons." *Political Theory* 28 (5): 640–674.

Nagel, T. 2005. "The Problem of Global Justice." *Philosophy and Public Affairs* 33: 113–147

Pettit, P. (2006). "Rawls's Peoples." In R. Martin and D. Reidy (eds.). *Rawls' Law of Peoples: A Realistic Utopia?* Malden: Blackwell.

Pogge, T. 1989. *Realising Rawls.* Ithaca: Cornell University Press.

Pogge, T. 2008. *World Poverty and Human Rights.* Cambridge: Polity Press.

Rawls, J. 1999a. *A Theory of Justice.* Cambridge: Harvard University Press.

Rawls, J. 1999b. *Collected Papers.* Cambridge: Harvard University Press.

Rawls, J. 1999c. *The Law of Peoples: with 'The Idea of Public Reason Revisited.'* Cambridge: Harvard University Press.

Rawls, J. 2000. *Lectures on the History of Moral Philosophy.* Cambridge: Harvard University Press.

Rawls, J. 2001. *Justice as Fairness: A Restatement.* Cambridge: The Belknap Press of Harvard University Press.

Richards, D. 1982. "International Distributive Justice." *Nomos* 24: 275–299.

Tan, K.-C. 1998. "Liberal Toleration in Rawls's Law of Peoples." *Ethics* 108 (2): 276–295.