Živan Lazović
Department of Philosophy
University of Belgrade

Original Scientific Paper
UDC

IS PUTNAM'S 'BRAIN IN A VAT' HYPOTHESIS SELF-REFUTING?

Abstract: *In this paper, I provide a detailed analysis of Putnam's conclusion (derived from the externalist interpretation of meaning and mental content) that the skeptical hypothesis, according to which we have always been brains in vats, is self-refuting. I confine my attention to the following question: If we assume that semantic externalism is plausible on independent grounds, does it follow that the semantic argument against skepticism (as articulated by Putnam) is indeed successful? In the first section, I briefly review the basic contention of Putnam's semantic externalism. In the second section, I outline and reexamine Putnam's, Brueckner's, and Warfield's version of the semantic argument. I hope to show that Putnam's version of this argument remains on a purely meta-linguistic level, which means that it can only prove that the phrase 'We are brains in a vat' must be false when it is considered in the context of the argument, although it most certainly does not prove that we are not brains in a vat after all. In the third section, I argue that Brueckner's and Warfield's attempt to modify Putnam's argument, and consequently provide an a priori proof that we are not brains in a vat, are ultimately unsuccessful, for both attempts beg the question against the skeptic. In the final section, I draw a comparison between the skeptical hypothesis and other cases of self-refuting statements and conclude that Putnam was ultimately right in claiming that the skeptical hypothesis is self-refuting in a weak sense, in which it is unassertible, although it might be true nevertheless.*

Keywords: *semantic externalism, reference, disquotation, self-knowledge, truth, assertibility.*

1. Introduction

In (1973, 1981), Putnam presents and articulates his externalist view of the meaning of referring (singular and general) terms, as well as of the content of our thoughts about physical objects. This view is widely known in the relevant philosophical literature as *semantic externalism* (SE). Although Putnam's primary intention in presenting his version of SE is to elucidate the nature of the relationship between our mind and the world, he indicates that SE can also be used to show that the skeptical hypothesis (SH)—in its most radical form, according to which we have always been brains in a vat (BIV) in an otherwise empty universe—is self-refuting (1981: 7). On the basis of these indications, later proponents of SE have offered the so-called *semantic*

argument (SA) against such a Cartesian-inspired form of skepticism. Given that both SE and SA have generated many philosophical discussions, I will not go into their detailed analysis here. Instead, I will focus on the following question: If we assume that SE is plausible on independent grounds, does it follow that SA (as articulated by Putnam) is indeed successful? In other words, the question I will be dealing with is whether SA shows that the skeptic's position is self-refuting in the sense in which, under the assumption that SE represents the correct view, it follows that SH must be false. I will argue in this paper that, given SE, the skeptic's position is in fact self-refuting in the weak sense, according to which SH is unassertible, although it might be true nevertheless.

2. Putnam's Semantic Externalism and BIV Hypothesis

Before presenting and analyzing SA, I will provide a brief explanation of SE. Thus, in Putnam's view, the basic thesis of SE could be formulated as follows:

> (SE): Reference and hence the meaning of referring terms (such as 'water', 'tree', 'table' and the like), as well as the content of our thoughts about the objects to which these terms refer, is at least partially determined by environmental factors, among which the causal link between the use of these terms and the objects to which they refer plays a prominent role.

In order to obtain a fuller understanding of how the causal constraint, expressed in the above formulation, determines the meaning of the referring terms and the content of sentences and thoughts that include them, three points are especially worth noting. First, the meaning of a referring term is at least partially determined by the direct causal link, established via its original introduction into the linguistic practice, as well as by the indirect causal chain of its later applications by the members of the language community. Thus, for instance, the term 'water' is first introduced by the speakers who have had direct causal encounters with individual samples of a substance with such-and-such properties. This direct causal link is then extended *via* the appropriate chain of communication, wherein the speakers use the term 'water' to speak and think about water.

Second, if it turns out that one and (linguistically) the same referring term has a different causal history in two language communities, then it will also have a different meaning in these language communities. This is shown by Putnam's famous 'Twin Earth' thought experiment (Putnam 1973), where we are told to imagine a scenario in which on a planet (nearly) identical to ours—the so-called *Twin Earth*—there are people who represent our physical and phenomenological duplicates and who, perhaps unsurprisingly, speak the

same language as we do. Now, both on Earth and on Twin Earth, there is a substance that Earthlings and Twin Earthlings identify by their superficial (i.e. observable), stereotypical properties as water and refer to its samples by applying the word 'water'. The only difference is that on Earth, the molecular structure of this substance is H_2O, while on Twin Earth, it is an entirely different substance with the molecular structure XYZ. Under the assumption that the molecular structure constitutes the identity of a substance, from the fact that Earth and Twin Earth differ with respect to the external environment and the causal history of the term 'water', it follows that this term has different meanings in English and vat-English. That is to say, when Earthlings use the word 'water', they refer to a substance composed of H_2O, whereas their duplicates on Twin Earth refer to a substance with the molecular structure XYZ. This fact brings about a difference with respect to the content of the appropriate sentences that we and our duplicates on Twin Earth formulate when we talk about water, as well as to the content of our thoughts about water, for the truth conditions of the sentences (and thoughts) about water on Earth differ from their truth conditions on Twin Earth. Namely, when we (on Earth) point to a sample of the liquid in a glass in front of us and say 'This is water', and when our duplicates on Twin Earth do the same, our sentence will be true if the liquid in the glass has the molecular structure of H_2O, while the sentence of our duplicates on Twin Earth will be true if the liquid in the glass in front of them has the molecular structure XYZ.

Third, if the speaker fails to meet the causal constraint on a referring term—that is, if she has never been in direct or indirect causal contact with the object to which she applies the referring term—then she cannot form the corresponding concept, make any assertions or, ultimately, have any thoughts *about* the object to which this term refers. We can thus see that the semantic significance of the corresponding causal link between the referring expressions and the objects to which they refer has a strong impact on the linguistic competence of the speaker. Namely, in order to be able to properly understand the meaning of a referring expression—that is, to use it correctly in speech and to think about the objects to which it refers —it is necessary to be in the appropriate direct or indirect causal contact with these objects. It is also important to note that, although Earthlings and Twin Earthlings use the term 'water' with different meanings, they can still successfully satisfy the causal constraint within their own language communities. Yet, someone—whether on Earth or Twin Earth—who has never been (directly or indirectly) in the appropriate causal contact with any sample of water, would be utterly unable to understand the meaning of the word 'water' or, consequently, formulate sentences and form thoughts about water (see, Putnam 1981: 12, 16; Kallestrup 2012: 36).

Now, as mentioned at the outset, Putnam was convinced that SE could serve as a powerful argumentative tool against the Cartesian-inspired philosophical skeptic. In order to avoid any possible quandaries about

whether we can meet the causal constraint in hypothetical situations in which we are only temporarily victims of the systematic deception, or in which we are constantly deceived by any other subject that otherwise meets the causal constraint, Putnam introduces (for the sake of argument) the most radically updated version of the Cartesian skeptical hypothesis:

> (SH) In an otherwise completely empty world, as a result of some cosmic accident, we are always disembodied brains envatted in a nutrient fluid, connected to a super-computer and having experiences, including thoughts, that are caused only by computer-generated electrical impulses.

It is worth pointing out that the scenario described in SH shows striking similarity to the above-mentioned Twin Earth thought experiment. Namely, observe that BIVs in SH should be understood as our phenomenological twins; that is, they represent our exact psychological duplicates with respect to sensory evidence, thoughts and interior monologue. In a BIVs' world in which there are no physical objects, the super-computer produces experiential experiences in the BIVs' minds that are qualitatively indistinguishable from the experiences that we have in our actual environment. We can thus see that in SH, the semantic point about the reference of the term 'water' from the Twin Earth thought experiment is extended to cover *all* referring terms in the BIVs' world. Suppose a glass containing liquid is in front of me and that, on the basis of my sensory evidence, I identify that liquid as water. Suppose further that I say, 'This is water', expressing with this sentence the content of my thought about the liquid in front of me. According to SH, it follows that my BIV—in its otherwise empty world—has the same sensory evidence produced by the appropriate electrochemical stimulation and that—in its own interior monologue—it utters the same sentence 'This is water'. Now, in this waterless world, the BIV cannot have any causal contact with water as a physical liquid, but rather with entities that in *its* own world play a causal role with respect to *its* uses of 'water' that is analogous to the causal role that the instances of water play with respect to *my* uses of 'water'. If these entities in the BIVs' world are, say, the recurring computer program features, then, according to the causal constraint of SE, the BIVs' word 'water' does not refer to water but rather to the recurring computer program feature <W>, which causes electrical stimuli in BIVs and, in turn, produces experiences that are qualitatively indistinguishable from the experiences of our embodied brains that are stimulated as a result of seeing water in normal circumstances.

The difference between the reference of the word 'water' in the actual world (in which we are normal human beings in our typical physical environment) and the reference of the word 'water' in the BIVs' world brings about an important difference in the semantic content (i.e. the truth-conditions) of my sentence 'This is water' and the BIV's sentence 'This is

water' respectively. Namely, in the actual world, my sentence will be true if the liquid in the glass in front of me really is water; in the BIVs' world, however, the sentence 'This is water' will be true if the computer program feature <W> is running. In other words, given SE, with the phrase 'This is water', my BIV and I (each in our own world) assert *different* statements and express *different* thoughts—of course, only if we assume, following Putnam, that BIVs can have any thoughts.

On the basis of this observation, it seems that the semantic difference between our and BIVs' sentences and thoughts can be successfully represented by using a disquotational mechanism, as a device that we use in ordinary (natural) language—given that it is semantically closed (i.e. contains semantic predicates which include both 'referring to' and 'true') and universal (i.e. contains both object– and meta-language)—in order to explicate both the reference of terms and the truth-conditions of declarative sentences.[1] Thus, by applying a disquotation mechanism, the reference of my word 'water' in English is determined by the following sentence:

(R_E) 'Water' refers to water;

and the truth-conditions of my sentence 'This is water' (which I assert while pointing to the liquid in a glass) is determined by the following equivalence:

(T_E) 'This is water' is true iff this is water.

Now, let us suppose that in my case, this truth-condition is obtained; i.e. that the liquid in a glass I am pointing to is, in fact, water. If we attach the same meaning to the words of BIVs—in their waterless and glassless world—the sentence 'This is water' would not be true. But given the causal constraint involved in SE, the word 'water' in vat-English does not refer to water, but rather to the computer program feature <W>. Since this is so, it follows that the truth-condition of BIV's sentence 'This is water' is obtained?? when the computer program feature <W> is running. In other words, at least from the perspective of our language—used as a meta-language for BIVs' object-language—it seems that the usual disquotation mechanism is not applicable in vat-English,[2] for we cannot obtain the reference to the word 'water' or the truth-condition of the sentence 'This is water' in vat-English by simply removing the quotation marks; rather, we would have to use the following formulations in *our own language*:

(R_E) 'Water' refers to a computer's program feature <W>;

that is,

(T_E) 'This is water' is true iff the computer's program feature <W> is running.

---

1 When sketching his argument in *Reason, Truth and History* (1981), Putnam does not use disquotation, but he resorts to it in his 'Replies' (1992: 347–408).

2 Cf. Brueckner 1992: 205.

Although the BIVs in SH are represented as our phenomenological duplicates, their language seems to lose the semantic properties of closedness and universality,[3] while the reference of their individual words and the truth-conditions of their sentences cease to have the disquotational character that is familiar to normal English speakers, due to the fact that the words they use in their empty environment are causally connected to the features of the computer program, and not to familiar physical objects.

Now, what are the consequences of all this for the hypothesis (SH), which the skeptic, as a normal speaker, wants to formulate in a natural language? Suppose this hypothesis simply reads 'We are brains in a vat'. By formulating this hypothesis in *our* language, the skeptic certainly takes the words 'brain' and 'vat' with their *usual* meaning, in which they refer to *brains* and *vats* as physical objects; that is, objects with which they are causally related. Given the formulation of SH, it is clear that the skeptic must not assume that *we have not always been* brains in a vat, or, ultimately, that *she herself is not*—at least at the moment in which she is presenting her skeptical hypothesis—a brain in a vat. But if SE is correct, the skeptic's position appears to be self-refuting for *semantic* reasons, that is, on the basis of the meaning of the terms used in the formulation of her hypothesis, as well as on the basis of its semantic content. This is exactly what Putnam claims. In his view, if SE is correct, the hypothesis that we have always been brains in a vat '*cannot possibly be true*, because it is, in a certain way, self-refuting' (1981: 7); that is to say, it represents a supposition whose truth implies its own falsity. Now, Putnam makes it clear that a statement can be self-refuting in at least two different ways (1981: 7–8). First, there are statements that are self-refuting only because of their semantic content, regardless of whether anyone asserts them or not; such is, for example, the statement 'All statements are false', for if this statement is true, it follows that it must be false. But there are also statements that are self-refuting partly due to their grammatic form; i.e. due to who and in which form asserts or contemplates them. Putnam's example of this particular type of statement is 'I do not exist', which, given the meaning of the pronoun 'I', must be false whenever (and in all circumstances in which) *any* person asserts or contemplates it in the present tense. After recalling this distinction, Putnam states that SH represents an instance of the second group of self-refuting statements:

> What I will show is that the supposition that we are brains in a vat has just this property. If we can consider whether it is true or false, then it is not true (I shall show). Hence it is not true. (1981: 8)

In the remainder of the paper, I will attempt to show that the fact that SH falls into the second group of self-refuting statements significantly diminishes the power of Putnam's version of SA against skepticism; namely, I will argue that

3 Ibid. 211.

from the fact that, given SE, neither the skeptic nor any of us can claim—in the usual sense in which 'brains' and 'vats' refer to *real* brains and *real* vats—that SH is true without implying that SH is false, it does not follow that SH cannot be true after all. In the next section, I will consider in more detail Putnam's versions of SA, but I will also pay attention to Brueckner's and Warfield's versions of this argument.

3. Putnam's Semantic Argument against the Skeptic

Putnam has tried to show that SH must be false by appealing to SE and its causal constraint on reference. As we have seen, according to SE, the words that BIVs use in their otherwise empty world, although linguistically the same as the words we use in the actual world, cannot refer to ordinary physical objects, given that BIVs have never been in causal contact with these objects. We have already agreed that in the BIVs' world, these words refer to the corresponding computer program features with which the tokens of their uses are causally connected. This also applies to the words 'brain' and 'vat' involved in SH. Thus, just as in the BIVs' world—where the word 'water' does not refer to water as a physical substance, but rather to the computer program features <Ws> with which BIVs' tokens of that word are causally connected—the words 'brain' and 'vat' in SH do not refer to *brains* and *vats* as physical objects, but rather to the computer program features <Bs> and <Vs>.

Following Brueckner (1986), we will present Putnam's argument in the disjunctive form, according to which we are either BIVs or we are not BIVs. The disjunctive formulation of this argument has the following consequences. If we are not BIVs, then by uttering the sentence 'We are BIVs' we mean that *we are BIVs* and, taken with this meaning, the sentence is clearly false. On the other hand, if we are BIVs, then by uttering the sentence 'We are BIVs' we would mean that we *are <Bs> in <Vs>*. However, since SH represents the hypothesis about *real* brains and *real* vats, rather than about the computer program features <Bs> and <Vs>, the BIV's sentence 'We are BIVs' turns out to be false. Given that the sentence 'We are not BIVs' is false whether or not we are BIVs, it follows that the opposite sentence 'We are not BIVs' must be true (cf. Putnam 1981: 14–15).

Yet, as Brueckner (1992) rightly observed, Putnam's argumentation works on a meta-linguistic level, which proves that the *sentence* 'We are BIVs' (when we assert or consider it) must be false. We therefore need at least one additional step in order to reach the conclusion that *we are not BIVs*. Proponents of SE typically maintain that this step could be made either by applying the disquotation mechanism (Putnam 1992; Wright 1992; Brueckner 1986, 1992) or, alternatively, by invoking the assumption—the so-called *self-knowledge thesis* (SK)—that the subject has privileged access to the contents of her mental states (Tymoczko 1989, Warfield 1998, Brueckner 2003). Let us consider these two strategies in turn.

Brueckner (1986) argued that we can reach the conclusion that we are not BIVs by applying the following disquotation principle:

(T) 'We are not BIVs' is true iff we are not BIVs.

Combined with the conclusion of Putnam's original SA, according to which the sentence 'We are not BIVs' must be true, the equivalence (T) leads us to the further conclusion that we are *not BIVs*. However, Brueckner himself (1986: 164–165) expressed concern that the application of the disquotation principle (T) in the context of disjunctive SA begs the question against the skeptic. As Folina (2016) and McKinsey (2018) show, this concern is well grounded. First of all, in the context of argumentation that starts with the disjunctive premise 'Either we are not BIVs or we are BIVs', (T) is most certainly *ambiguous*. Namely, note that whether we speak normal English or vat-English depends on whether or not we are BIVs. In either of these two languages—i.e. as (T_E) or as (T_{VE})—the equivalence is the same: 'We are not BIVs' is true iff *we are not BIVs*. But the truth-conditions for the above-mentioned sentence 'We are not BIVs' are evidently different in these two languages. Thus, if we are not BIVs (i.e. if we speak normal English), the truth-conditions are that *we are not BIVs*. If, on the other hand, we are BIVs (i.e. if we speak vat-English), the truth-conditions are that *we are not <Bs> in <Vs>*. The conclusion in the vat-English sense that *we are not <Bs> in <Vs>* obviously misses the point, since we wanted to get to the conclusion that we are not BIVs in the normal English sense. However, in order to reach this particular conclusion, we would have to employ (T) within normal English (with the subscript $_E$), but given the starting disjunctive premise of SA, it turns out that to assume that we are normal English speakers is to assume in advance the point we wanted to prove; namely, that we are not BIVs (cf. Brueckner 2016: 4; Kallestrup 2018: 170).

In his later reconstruction of SA, Putnam applied a disquotation scheme (R) by arguing that from the fact that our word 'water' refers to *water*, with whose instances we are causally connected, it follows that we are not BIVs in the waterless world (1992: 369). Brueckner's (2003) simple version of this argument runs as follows:

(Br1) If we are BIVs, then our word 'water' does not refer to water.
(Br2) Our word 'water' refers to water.
(Br3) So, we are not BIVs.

However, step (Br2) is controversial for two important reasons. First, in order to know that our word 'water' refers to *water* and not to *<W>*, we would have to *know* that our uses of this word are indeed causally linked to the instances of water as a liquid in our normal physical environment, where this knowledge must be *empirical* and, as such, endangered by SH. Second, our uses of the word 'water' refer to *water* only if we are normal English speakers in a normal physical environment, and since this can only be the case if we are not BIVs, we seem to beg the question against the skeptic once again.

Arguably, within a semantically closed language, we use disquotation as a syntactic means by which we present the reference of the terms and the truth-conditions of the sentences containing these terms. The knowledge that we—as normal competent speakers—possess about the role of quotation marks, as well as of the semantic terms 'refers' and 'true', is indeed a priori in that it allows us to present the reference of any meaningful referring term '*m*' with the scheme (R): "'*m*' refers to *m*", and the truth-conditions of any sentence '*s*' with the equivalence (T): "'*s*' is true iff *s*". However, the lesson from the Twin Earth thought experiment is that without additional descriptive information about the objects with which our uses of words are causally connected, disquotation is utterly insufficient to determine the reference of expressions or the truth-conditions of the sentences.

Even before the discovery of the molecular structure of liquids, Earthlings and Twin Earthlings could—each in their own language—successfully apply (R_E or R_{TE}): "'water' refers to water", and (T_E or T_{TE}): "'This is water' iff this is water". Namely, before the discovery of the difference in the molecular structure of that liquid on Earth and on Twin Earth, we were willing to argue that the word 'water' both in Earth English and in Twin Earth English has the same reference, and that sentences about water have the same truth-conditions. However, it is worth pointing out that the meaning of the word 'water' and the truth-conditions of the sentences about water in Earth English and Twin Earth English did not become different the moment we came to this discovery (see Putnam 1973: 702). Namely, given SE, it is clear that the uses of the word 'water' in Earth English and Twin Earth English had different references even before that discovery and that the utterances of the corresponding sentences had different truth-conditions all along. Of course, we were not in a position to detect this difference by applying (R) and (T), which read the same in both languages, but only through empirical research. Therefore, in order to specify this difference, it is necessary to supplement the right side of (R) and (T) with the appropriate descriptions: in (R_E) 'water' refers to water as liquid H_2O, and in (R_{TE}) 'water' refers to water as liquid XYZ; also, in (T_E) 'This is water' is true if this liquid is H_2O, and in (T_{TE}) 'This is water' is true iff this liquid is XYZ.

Now, the same lesson applies to the reference of the word 'water' in the context of Brueckner's simple version of SA. One, perhaps not so important, difference with the Twin Earth experiment is that there is no water in the BIVs' world and that the most suitable candidates for external causes of BIVs' uses of the word 'water' are the computer program features <Ws>. Hence, according to SE, all BIVs' uses of the word 'water' refer to <Ws>. But, from the perspective of both normal English and vat-English, the application of a disquotation (R) to the word 'water' provides the same linguistic outcome: "'water' refers to water". As such, in order to express the semantic difference between our and BIVs' uses of the word 'water', we need to supplement the right side with an adequate descriptive characterization that distinguishes

the objects with which these uses are causally connected in our and BIVs' environments. That is to say, the word 'water' in our environment refers to the instances of such-and-such physical liquid, and in the BIVs' environment, it refers to the computer program feature <W>. The limitations of this strategy are by now more than obvious. Namely, the main difficulty here arises from our utter inability to know what kind of environment we *de facto* inhabit, that is, from the fact that we can never know whether we are normal English-speaking human beings in an environment with physical objects, or whether we are BIVs in a completely empty environment.

Now, let us see if the observation that vat-English is not semantically closed and that the reference and the truth-conditions in it are not disquotational is of any help. It is precisely on this observation that Brueckner (1992) articulates one version of his SA, but he *relativizes* it with respect to normal English as a meta-language. Since we know in advance from SH that BIVs' uses of the word 'water' do not have the same reference as *our* uses of that word, and that the truth-conditions of the BIVs' sentences 'This is water' are not the same as the truth-conditions of *our* uses of that sentence, by applying the disquotation schemes (R) and (T) in our English to BIVs' use of the word 'water' and to their sentence 'This is water' we will not get accurate results; on the right side of the disquotational schemes (R_E) and (T_E), we need to put <*W*> and *this is* <*W*> instead of the *water* and *this is water* mentioned on the left side. But does this mean that vat-English is not semantically closed *tout court* and that the reference of words and the truth-conditions of the sentences in this language are not disquotational *independently* of our English? In order to provide a satisfactory answer to this question, I think it is instructive to appeal once again to the Twin Earth thought experiment: if we have no reason to question that Twin Earth English is semantically closed and that disquotation works within this language, then there seems to be no reason whatsoever to question that the same is the case with vat-English.

Similar to the impression that we had, following Brueckner, with respect to the semantic difference between our and BIVs' sentences, when we realize that the word 'water' in Twin Earth English refers to the instances of XYZ, at first glance it might seem to us (from the perspective of Earth English as a meta-language) that in Twin Earth English neither the reference of that word nor the truth-conditions of the corresponding sentences are disquotational, for we should represent them as follows: in Twin Earth language, 'water' refers to *XYZ*, and 'This is water' is true iff *this is XYZ*. However, just like Earth English, Twin Earth English has all the linguistic resources that make it semantically closed and allows for disquotation: it contains the semantic terms 'refer' and 'true' as well as quotation marks, as a syntactic means by which we name linguistic expressions.

The difference in the reference of the word 'water' and the truth-conditions of the sentence 'This is water' in normal English and vat-

English have an empirical rather than a linguistic origin: the application of disquotation in both languages yields the same outcome, but the important difference between our and the Twin Earth environment consists in the fact that the named liquid on Earth is H_2O, whereas on Twin Earth it is XYZ. Due to this empirical discovery, we can specify the reference and the truth-conditions by adding the appropriate descriptive characterization on the right side of the disquotation schemes. There is no reason why inhabitants of Earth and Twin Earth should not continue to use the word 'water', as well as the sentence 'This is water', in the same way as they used them before this discovery, and why they could not (each in their own language) successfully apply the disquotation, while at the same time being aware that in their two languages (owing to the difference with respect to their environment) the word 'water' has different references and the sentence 'This is water' has different truth-conditions.

As for the semantic closeness and the applicability of disquotation, it seems that what is true of Twin Earth English is also true of vat-English. According to SH, vat-English (just like Twin Earth English) has all the linguistic resources that make it semantically closed and allows for disquotation: it contains the semantic terms 'refer' and 'true' as well as quotation marks, as a syntactic means by which we name linguistic expressions. The difference with respect to the reference of the word 'water' and the truth-conditions of the sentence 'This is water' in normal English and vat-English has an empirical rather than a linguistic origin: the use of disquotation gives us the same outcome again, but our and BIVs' environments differ in that the uses of the word 'water' in our language sustain a causal connection to the instances of such-and-such physical liquid, whereas in the BIVs' waterless world, it sustains a causal connection with the computer program features <Ws>. The only important, and yet empirical, difference between the Twin Earth and the BIVs' world is that BIVs (unlike the Twin Earthlings) are utterly unable to discover to which particular objects in the environment the word 'water' is causally connected; that is, they are unable to find an adequate descriptive characterization to determine the reference and truth-conditions of their linguistic expressions and sentences. It is especially worth stressing, however, that the semantic closedness of a language, as well as the possibility of applying disquotation in it, should depend only on the linguistic resources, and not on the epistemic position of the speakers. In other words, it is irrelevant for these linguistic features whether we *sometimes* make errors in descriptively identifying objects of reference (as in the Twin Earth experiment), or whether we *always* and *systematically* make such errors (as in the BIVs' world). Thus, in contrast to the Twin Earth thought experiment, the main point of the skeptical BIV hypothesis is that we might be in the BIVs' position after all. If we are not able to exclude this possibility, we will never know for certain that our uses of the word 'water' refer to the instances of such-and-such liquid. Given SE, the most we can know is that the following conditionals "If we are in a

normal environment, 'water' refers to the instances of such-and-such liquid" and "If we are BIVs, 'water' refers to the computer program features <Ws>" are true. Unfortunately, we cannot know—either a priori (i.e. by means of disquotation) or a posteriori (i.e. by means of sensory evidence)—which of the two antecedents is in fact true; that is, we cannot know whether we are normal human beings speaking normal English or BIVs speaking vat-English.[4]

We have thus seen that disquotation cannot help us to complete Putnam's SA and, consequently, prove that we are not BIVs. Some semantic externalists (e.g. Tymoczko 1989, Warfield 1998, Brueckner 2003) have used the fact that SE represents the thesis about the content of our thoughts of external objects, and appealed to the assumption that was traditionally considered to be an internalist ally: with respect to the contents of our thoughts, we have immediate, privileged access that allows us to obtain non-evidential, a priori self-knowledge (SK) about *the content* of our thoughts. One version of SA that relies on SK was offered by Warfield (1998: 78):

(Wr1) I think that this is water.
(Wr2) In its waterless world, no BIV can think that this is water.
(Wr3) So, I am not BIV.

As we can see, the first premise of Warfield's argument relies on SK, and the second on SE. Of course, this particular combination of premises can only be legitimate if these two theses are compatible. Yet, as is well known, Putnam considered SK to be inconsistent with SE,[5] and his opinion was shared by many authors (e.g. McKensey 1991, Bilgrami 1992, Brown 1995, Boghossian 1997, etc.). Some of them, such as Bilgrami (1992), argued that, if SE (along

4 The same line of reasoning can be applied in order to refute some recent attempts (e.g. Thorpe 2018) to prove that we are not BIVs on the basis of the assumption that the subject has non-empirical semantic knowledge of the content of his current thoughts and that this knowledge can be expressed in the disquotational form (e.g. "My thought 'This is water' has the content that this is water"). Given the limitation of space, I cannot provide a detailed analysis of this proposal. See Falvey & Owens (1994) for the point that we cannot have non-empirical knowledge of the comparative content of our thoughts: namely, in order to find out that the content of my thought 'This is water' would be different depending on whether I form this thought in the Earth or Twin Earth environment, I would have to find out the difference in the molecular structure of water, which is something that I can only know empirically.

5 Putnam found a solution to the conflict between SE and SK by bifurcating the content of thoughts into narrower and wider. The thought of an object in its narrowest content is the subject's conception of the object as an internal psychological state, while in its wider content, this thought is determined by the external relation to the object. Thus, in the Twin Earth thought experiment, when we have a thought expressed by the phrase 'This is water', it turns out that we and our phenomenological duplicates share the same *narrow* thought content (i.e. we are in the same psychological state when the instance of the substance we perceive is called 'water'), but we have different *wide* thought contents (i.e. our thought is *about* a sample of the liquid H_2O, and our Twin Earth duplicate's thought is *about* a sample of the liquid XYZ.).

with its causal constraint) is correct, then in order to possess knowledge of the content of thoughts (as well as to determine their references and truth-conditions), we would have to include a descriptive characterization of objects from the environment with which these thoughts sustain a causal connection. Yet, Bilgrami continues, we cannot know that such characterizations are accurate without a proper (a posteriori) empirical investigation. Others, such as McKinsey (1991), argued that the combination of SE and SK leads to absurdity: according to SK, we should have privileged access to our thoughts and thus know a priori that we are thinking a water-thought; on the basis of SE, on the other hand, we should know a priori that if we have water-thoughts, then water exists. From these two premises, it follows that we should know a priori that water exists. However, given that we do not have privileged access to the outside world, our knowledge of the existence of water must be a posteriori (cf. Kallestrup 2012: 173).[6]

In the light of many discussions on this topic, the ultimate impression is that the incompatibility of SE and SK cannot be eliminated without rejecting or, at least, significantly modifying one of these theses. Thus, for example, Bilgrami (1992) modifies SE by introducing a fundamentally internalist constraint, according to which, in selecting the object in the environment that is supposed to fix the concept that is being expressed by the given term, one has not only to pick the object which is obviously causally correlated with that term but also to *describe* this external determinant of the concept 'in a way that fits in with the other *contents* one has attributed to the agent' (1992: 257). On the other hand, Nuccetelli (2003) modifies SK by distinguishing between two types of a priori knowledge: in the first, stronger sense, the knowledge of a statement is a priori if it is completely independent of empirical assumptions, while in the second, weaker sense, the knowledge of a statement may be a priori even if it includes certain empirical assumptions in light of which that statement can be challenged a posteriori; in her view, self-knowledge about the content of the thought 'This is water' is a priori in the weaker sense, for, according to SE, it rests on the empirical assumption that the term 'water' does have a reference with which it is causally connected—namely, it refers to the instances of H_2O (2003: 180).

Either way, it turns out that neither disquotation nor SK can help us in completing Putnam's SA. If we accept the original (Putnam's) version of SE, we are forced to reject or at least significantly modify SK: without the additional evidential knowledge of the environment and the objects to which our thoughts are causally connected, we cannot fully know the contents

6 The thesis that Putnam's versions of SE and SK are incompatible could be defended in another, indirect way. Namely, under the essentialist assumption—according to which the molecular structure essentially determines the identity of substances such as water—someone who does not know the molecular structure of water could entertain the thought 'Water is not H_2O'. If both SE and SK are accepted, this thought should be logically inconsistent. Hence, if we want to preserve SE, without declaring that person irrational, we have no other option but to reject SK (Bilgrami 1992).

of these thoughts. The premise (Wr1) in Warfield's argument is therefore problematic and questionable for similar reasons to the premise (Br2) in Brueckner's argument. Namely, in order to know that our thought of water is, indeed, the thought of water as a physical liquid (i.e. H_2O), and not of the computer program features <Ws>, we would have to know that our thought sustains the appropriate causal connection with the instances of H_2O. We cannot gain such knowledge a priori, but only a posteriori; that is, only through empirical research and on the basis of the appropriate sensory evidence. By formulating SH, however, the skeptic eliminates our possibility of having such evidence: on whatever sensory evidence we base our belief that our thought about water sustains a causal connection to the instances of H_2O, we cannot know for sure that this thought does sustain such a causal connection, for we cannot rule out the possibility that we are BIVs who have the same sensory evidence and the same (though false) beliefs about our environment. That is to say, just as we are convinced that we have a thought of water as a physical liquid, BIVs can be convinced (though wrongly) that they have a thought of water as a physical liquid. If, despite our inability to know this, we endorse (Wr1)—thereby implying that we have the thought of water as an instance of H_2O—we assume in advance that we are in a normal physical environment and, ultimately, beg the question against the skeptic.

4. Self-Refuting Character of the BIV Hypothesis

As we have seen in the previous section, both Brueckner's and Warfield's externalist attempts to complete Putnam's SA and, consequently, reach the conclusion that we are not BIVs have failed. Since there does not seem to be any third way to accomplish these goals, the ultimate reach of Putnam's argument against skepticism is the meta-linguistic conclusion that SH in the form of the sentence 'We are BIVs'—when it is claimed or considered (as is obviously the case in the context of the skeptical argument)—must be false. Now, where does this leave the skeptic?

Assuming that SE is correct and that SH is formulated as Putnam proposes, it turns out that in making his argument, the skeptic is forced into a somewhat precarious position. Namely, by presenting the possibility of a mistake that we seemingly cannot exclude, the skeptic must take the statement 'We are BIVs' in the normal English sense; that is, in the sense in which the words 'brain' and 'vat' refer to actual *brains* and *vats*. As such, the skeptic implies that we are normal English speakers (i.e. that we are not BIVs), thereby acknowledging the point of Brueckner's premise (Br2), as well as of the premise (Wr1) in Warfield's version of SA. If the skeptic starts with the assumption that the 'We are BIVs' hypothesis is true, she undermines the possibility of asserting this hypothesis in the sense she originally intended. In this case, the skeptic herself would be BIV without any causal contact

with real brains and real vats; in other words, she would be in a position in which she could only assert the sentence 'We are BIVs' in vat-English, and her words 'brain' and 'vat' would only refer to the computer program features <Bs> and <Vs>. Putnam's disjunctive version of SA points out this skeptic's predicament: whether or not we are BIVs, the 'We are BIVs' hypothesis is shown to be false.

Still, it is important to bear in mind the following restriction that Putnam himself invokes: *whenever we are considering* whether SH is true or false, it follows that it must be false. This restriction creates theoretical space within which the radical skeptic can find at least some sort of escape route. Namely, Putnam's SA proves only the unassertiveness (but not falsehood) of SH: expressed by the sentence 'We are BIVs', SH cannot be truly asserted by us, although the proposition expressed by this sentence might be true nevertheless. Put otherwise, the proposition expressed by SH is in itself perfectly consistent, but when we assert it, we contradict ourselves. Why is this?

Let us once again recall Putnam's observation about the self-refuting character of SH. He reminds us that there are two groups of self-refuting statements (Putnam 1981: 7–8). The first group includes statements that are self-refuting on the basis of their semantic content, regardless of whether anyone asserts them; the statement that falls into this group is the general statement 'All statements are false' which, if true, must be false. As is well known, in a semantically closed and universal language such statements give rise to semantic paradoxes (such as those of the Liar family), for they depend upon the semantic notion of truth and on explicit self-reference (i.e. the sentence refers to itself). In other words, they are self-refuting only due to their semantic content and self-reference, regardless of *linguistic* factors (for instance, the presence of terms that indicate *who* and in *what* circumstances *asserts* them) or *theoretical* factors (such as specific assumptions about the meaning of some terms that occur in a statement). As an example of self-refuting statements that fall into the second group, Putnam cites the statement 'I do not exist'. This sentence is understood to be self-refuting due to its semantic content. But since the indexical term 'I' introduces an element of self-reference, it is obvious that its self-refuting status depends also on *who* and in *what* linguistic form *asserts* the proposition expressed by this sentence: the statement must be false only if it is asserted by the speaker (or speakers) in the first person, present tense. Yet, note that the same proposition *about my non-existence* may be truly asserted by someone else (e.g. 'He (Živan Lazović) does not exist'), or even by me in some other (past or future) tense (e.g. 'I did not exist' or 'I will not exist').

So, despite the fact that the sentence 'I do not exist' is necessarily false when I assert it in the present tense, it does not follow that the *proposition* expressed by this sentence cannot be true after all. The first type of statements, whose self-refuting character depends solely on their semantic content, will

be characterized as self-refuting in the *strong* sense, while the second type of statements, whose self-refuting status depends partly on semantic content and partly on linguistic form (i.e. presence of particular expressions that introduce self-references by pointing out *who* and in *what* circumstances asserts them) or on some specific theoretical assumptions (such as the SE thesis), will be characterized as self-refuting in the *weak* sense. I think we should concede Putnam's claim that SH—presented in the form of the statement 'We are BIVs'—is self-refuting in the weak sense.

I will show in the remainder of this paper that the self-refuting character of SH rests partly on its semantic content (which is conditioned by the SE assumption about the meaning of the word 'brain' and 'vat') and partly on the fact that it is asserted in the first person form. I will also show that SH cannot be asserted by *any* human being, but that the proposition expressed by it might be true nevertheless. In this respect, the weak self-refuting status of the statement 'We are BIVs' is no exception. Without going into the exhaustive analysis, I will compare this statement with similar statements whose self-refuting character is partly dependent on their semantic content and partly on additional theoretical (conceptual) assumptions or the use of certain linguistic terms that make them unassertible relative to some particular speaker. Each of the following examples will be accompanied by brief remarks that should account for the fact that those who make such sentences contradict themselves without uttering a contradiction and, consequently, help us clarify the unassertibility of SH.

As an example of semantic paradoxes, the first sentence belongs to the well-known Liar family:

(1) 'All humans are lying.'

It is clear why, in a semantically closed and universal language, this sentence is self-refuting: it depends on the semantic notion of truth, the element of self-reference provides the universal quantifier 'all', and the sentence is unassertible relative to any speaker who belongs to the class—i.e. the class of human beings—about which it talks. Note, however, that this sentence might be consistently and truly asserted by any non-human being, as well as that the proposition expressed by this sentence might be true even if it is not asserted by anyone.

The so-called *Moorean paradox* provides us with yet another interesting example:

(2) '*p*, but I do not believe that *p*.'

Although I can assert *p* at some particular moment and add that I do not believe that *p* at some other moment, it seems that if I simultaneously say '*p*' and 'I do not believe that *p*', I contradict myself. Admittedly, there are numerous interpretations of this paradox in the relevant literature, but according to one of the most popular accounts, assertion and belief are

directed to truth in such a way that to assert *p* is to *express* or *imply* the belief that *p* is true. Hence, whoever asserts *p* and conjoins it with the assertion that she does not believe that *p*, obviously contradicts herself in the sense that she believes that *p* and she does not believe that *p*. It is worth stressing, however, that the self-refuting status of this sentence also depends on its formulation in the first person, present tense. Even if it is true that one would not assert that *p* without believing that *p*, one's belief that *p* does not imply *p*. Given that this is so, no problem will occur with the second and third person counterparts of (2)—e.g. '*p*, but *you* do not (*she* does not) believe that *p*'—or with the sentence in the first person past tense, such as '*p*, but I did not believe that *p*'. So, in spite of being unassertible in the first-person present tense, the Moorean sentence expresses a consistent proposition which might be true. Put differently, it may well be that *p*, and—just like in the example with 'I do not exist'—it might be true *about* me (or us) that I (or we) do not believe that *p*.

The third and even subtler example—closely related to some responses to skepticism—is the so-called *abominable conjunction* (see DeRose 1995):

(3) 'I know that I have hands, but I do not know that I am not BIV.'

It seems that even this sentence cannot be asserted in the first person without falling into contradiction. This is so because of the conceptual connection between knowledge and truth, and because of the fact that the statement 'I have hands' implies that I am not a (handless) BIV. This is analogous to the Moorean paradox in all important respects. Namely, if I claim to know that I have hands in the first conjunct, I thereby imply that I am not BIV, whereas in the second conjunct, I explicitly question this implication by allowing the possibility that I am BIV. The inconsistency becomes even more obvious if we assume the principle of deductive closure: from my knowing that I have hands and my knowing that having hands implies that I am not BIV, it should follow that I also know that I am not BIV. Thus, if I am willing to acknowledge that I do not know the implied statement, it follows by modus tollens that I must refrain from claiming to know the antecedent. This point is especially important given that the Cartesian versions of the skeptical argument—including the Putnamian BIV version of this argument—seemingly rest on the principle of deductive closure.

The self-refuting character of this conjunction, however, depends on the particular conception of knowledge that we assume. Thus, for invariantists and infallibilists—e.g. Descartes and Peter Unger respectively—this conjunction is self-refuting in the strongest sense, for both the first and the third person formulation (e.g. 'She knows she has hands, but she does not know that she is not BIV') sound like a contradiction in the same way as 'I know (She knows) that *p*, but it is possible that not-*p*'. Yet, for someone who is a fallibilist, and especially for those who (like Dretske and Nozick) reject the deductive closure principle, the third person formulation of the

conjunction will be perfectly consistent. However, it seems that even for these authors, there is a problem with the first-person version of the conjunction 'I know that I have hands, but I do not know that I am not BIV', for it looks self-refuting for a similar reason to the Moorean paradox: the problem is that the speaker negates the point implied by her first conjunct by asserting the second conjunct.

Mark Heller (1999) has shown that the abominable conjunction could be interpreted as self-refuting in the weak sense within a variantist conception of knowledge such as conversational contextualism. As is well known, the central thesis of conversational contextualism is that the concept of knowledge is context-sensitive in the sense that knowledge attributions of the form '*S* knows that *p*' can express different propositions (and thus have different truth-values) depending on the attributor's conversational context. These changes occur because knowledge attributions in different contexts can apply different—i.e. lower or higher—standards for knowledge. Contextualists explain the change in epistemic standards mainly by relativizing Drecke's idea of relevant alternatives with respect to the knowledge attributors: when we evaluate whether someone in a given context knows some statement *p*, we expect that person to exclude (*ceteris paribus*) all those alternatives (i.e. possibilities of error) which we consider relevant in this context. The contextual change of conversational factors—such as intentions, needs or interests of the knowledge attributors—results in the narrowing or widening of the set of relevant alternatives; that is, it results in lowering or raising the standard of knowledge. According to most contextualists, including Heller, in order to make an alternative *relevant* to the assessment of knowledge in a given context, it is sufficient to pay attention to it (Lewis 1996) or to make it salient (Cohen 1988). Thus, for instance, in everyday contexts of knowledge attribution, the epistemic standards are relatively low, which means that remote alternatives—e.g. various skeptical possibilities of error—are not taken into account. These skeptical possibilities of error, however, become relevant in the philosophical context. As such, our knowledge attributions in everyday contexts will (*ceteris paribus*) express truth, whereas in the philosophical context, they will express falsehoods due to the fact that skeptical hypotheses make salient precisely those alternatives which we are unable to exclude.

It is worth noting that, since Heller—like other conversational contextualists—has in mind knowledge attributions from the third-person perspective, the assertion of the abominable conjunction should be self-refuting regardless of whether we (as speakers) are attributing (or denying) knowledge to someone else or to ourselves. What is important is that, if the contextualist explanation is correct, we will be able to truthfully claim in ordinary contexts that we know (*ceteris paribus*) that we have hands in spite of not knowing that we are not BIVs. On the other hand, any emphasis of the possibility that we are BIVs would shift us into the skeptical context, in

which—given that we cannot know that we are not BIVs—it will not be true to know that we have hands. The resemblance to the Moorean paradox is now apparent. Each part of conjunction (3) can be asserted independently (of course, in different contexts). Yet, by asserting them simultaneously, we fall into contradiction: if we ascribe to ourselves (or to someone else) the knowledge of having hands, and at the same time maintain that we do not know that we are not (handless) BIVs, we make this skeptical alternative relevant and create a skeptical context, wherein no one can know that she has hands, because no one can exclude the possibility of being a (handless) BIV. As Heller concludes, this point 'explains the abominableness of DeRose's abominable conjunction' and 'makes the conjunction true but unassertible' (Heller 1999: 204). Thus, according to this explanation, the conjunction is true in everyday contexts of knowledge attribution in which *no one asserts it*, but it is unassertible relative to the participants in conversational contexts, for its assertion by any speaker in any context would turn the given context into a skeptical one and, thereby, make it false.

It is easy to see, I think, that all three examples considered above express consistent propositions in the same way as Putnam's example 'I do not exist', and become false only when they are asserted by some particular speaker; that is, when they are asserted in the first person, by the members of the class about which the proposition states something, or by the participants in the conversational context. The propositions expressed by these sentences are, therefore, unassertible relative to particular speakers, although they can be true nevertheless. Given that the same point applies to SH, it is clear that this hypothesis falls into the group of self-refuting statements in the weak sense. I will explain why this is so by drawing a comparison between SH and each of the three examples stated above.

The self-refuting character of Putnam's SH is certainly influenced—as was the case with (1)—by the element of self-reference invoked by the first person, present tense formulation; recall that such self-reference was also observed in the example 'I do not exist'. With respect to the version of SA that invokes the self-knowledge (SK) thesis, it is imperative that the argument is formulated in the first-person by using the indexical expression 'I' (see Wright 1992: 76–7; Kallestrup 2012: 172–3). In the context of Putnam's version of SA, however, the first-person formulation is not mandatory. By formulating SH, the skeptic addresses us as human beings and presents us with the possibility that we are BIV, which points to a fatal flaw in our epistemic position and, ultimately, compromises our knowledge of the external world. The skeptic can therefore formulate SH in the form of the universal statement 'All humans are BIVs' in which—given that the property of being BIV is ascribed to all members of that class—self-reference occurs if the statement is asserted by any member of the class of human beings (this is yet another similarity to (1)). In this formulation, it is clear that the statement is unassertive relative to *us* as human beings. As such, some other (non-human) being could assert

this sentence without any problems, and the proposition expressed by this sentence could be true even if no being asserts it.[7] Putnam's SH is, therefore, self-refuting when it is asserted in the first person, or when it is asserted in the form of a universal statement by any speaker who belongs to the class of human beings, since in both cases it contains self-reference.

We have seen that in (2) the speaker finds herself in a paradoxical situation, for by asserting the second conjunct, she negates what is implied by the first conjunct. By asserting SH in either of the two mentioned forms, the speaker makes the statement false by undermining a necessary condition—i.e. the existence of a proper causal link—under which her words 'brain' and 'vat' can have the desired meaning in the context of considering the skeptical argument; i.e. she makes it impossible for her words to refer to *real* brains and *real* vats. Something similar occurs in the example 'I (we) do not exist'. Namely, when uttering a statement, it is a necessary condition for the proper use of the pronoun in the first-person present tense that the speaker exists at the moment of utterance. So, by denying her own existence in the second part of the sentence, the speaker denies the fulfillment of that necessary condition and thus falls into contradiction.

Finally, SH has in common with (3) that a particular theoretical assumption concerning the meanings of the used terms is crucial for its unassertiveness. In the case of (3), it is a contextualist assumption about the meaning of the concept of knowledge, whereas in the case of SH, it is an externalist assumption about the meaning of referring terms such as 'brain' and 'vat'. We have seen that the (externalist) causal constraint on the reference in the context of Putnam's disjunctive SA leads to the conclusion that the words 'brain' and vat' have different references depending on whether or not we are BIVs: in the second case, these words refer to brains and vats as physical objects, while in the first they refer to the computer program features <*Bs*> and <*Vs*>. We have also seen that in order to derive her conclusion, the skeptic is forced to demonstrate that SH is true in the normal English sense. But due to the causal constraint on the reference and the element of self-reference in both formulations of SH, it follows that the skeptic—when considering whether SH is true either in the first person or as a member of the class SH refers to—finds herself in a paradoxical position. Namely, it turns out that when the skeptic asserts 'We (all humans) are BIVs', she does not assert a true proposition in the normal English sense (and does not mean that we are *real* brains in *real* vats), but rather a quite different, and ultimately false, proposition that we (humans) are <Bs> and <Vs>.

7 Although SH is construed in such a way to include the assumption that, aside from BIVs, the actual world does not contain any other intelligent beings that could claim that *we* (i.e. human beings) are BIVs, it does not mean that some (logically) possible world in which such intelligent beings exist, and in which they could truly assert the proposition 'All humans are BIVs', is inconceivable. I am strongly inclined to think that this is the status of SH that Crispin Wright had in mind when he made his remark that Putnamian SA does not refute the skeptical nightmare (1992: 73, 93); cf. Kallestrup (2012: 173).

5. Conclusion

On the basis of previous considerations, I think it is safe to say that Putnam was essentially right in claiming that the BIV hypothesis is self-refuting in the weaker sense. However, such self-refuting status has been shown to imply merely the unassertiveness of this hypothesis and is only relative to *us* as normal (human) speakers. In other words, Putnam's SA only proves that the sentence or statement 'We are BIVs' must be false when it is *us* (humans) who question its truthfulness. For this reason, SA has limited reach against the skeptic. The same consistent proposition that is expressed by the sentence 'We are BIVs' can be expressed in some other grammatic form—e.g. 'All humans are BIVs' or 'You are BIVs'—and it may be truthfully asserted or considered by some other (non-human) being, and can even be true when it is not asserted.

In the end, I should clarify that my intention in this article was not to provide a philosophical defense of the skeptic's position. Yet, when faced with the inability to consistently think or assert the truth of SH, the skeptic can do one of the following things. First, she can appeal to some version of the skeptical hypothesis that is completely beyond the reach of Putnamian SA. Second, she can acknowledge Putnam's SH and find at least some sort of satisfaction in the fact that it is still possible for this hypothesis to be true. Given that Putnam's SA was not successfully completed by any of its subsequent versions, the main point of the skeptical argument persists: it is indeed possible that we are BIVs, and there does not seem to be any theoretical or empirical strategy to exclude this possibility. Unfortunately, I have to admit that the same conclusion seemingly applies to the thoughts expressed in this paper. For, despite the fact that I cannot truthfully say or consistently think that *I am* (and have always been) a BIV, the proposition 'Živan Lazović is a BIV' may be true after all. But if this proposition is true, and if SE represents the correct view, then the words and thoughts expressed in this paper ultimately concern computer program features instead of real objects in the external world.[8]

References

Bilgrami, A. 1992. Can Externalism Be Reconciled with Self-Knowledge? *Philosophical Topics*, 20: 233–268.

Boghossian, P. 1997. What the Externalist Can Know A Priori. *Proceedings of the Aristotelian Society*, 97: 161–175.

Brown, J. 1995. The Incompatibility of Anti-individualism and Privileged Access. *Analysis*, 55: 149–156.

8 I wish to thank the anonymous referees for their very helpful comments on earlier drafts.

Brueckner, A. 1986. Brains in a Vat. *Journal of Philosophy*, 83: 148–167.

–, 1992. Semantic Answers to Skepticism. *Pacific Philosophical Quarterly*, 73: 200–219.

–, 2003. Trees, Computer Program Features, and Skeptical Hypotheses. In Stern, R. (ed.), *Transcendental Arguments: Problems and Prospects*, Oxford: Clarendon Press, pp. 152–162.

–, 2016. Skepticism and Content Externalism, *Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2016/entries/skepticism-content-externalism/

Cohen, S. 1988. How to Be a Fallibilist. *Philosophical Perspectives*, 2: 91–123.

DeRose, K. 1995. Solving the Skeptical Problem. *The Philosophical Review* 104: 1–52.

Dretske, F. 1970. Epistemic Operators. *The Journal of Philosophy*, 67: 1007–1023

–, 1981. The Pragmatic Dimension of Knowledge. *Philosophical Studies,* 40: 363–378.

Falvey, K., & Owens, J. 1994. Externalism, Self knowledge and Scepticism. *The Philosophical Review*, 103: 107–137.

Folina, J. 2016. Realism, Skepticism, and the Brain in a Vat. In Goldberg, S. C. (ed.), *The Brain in a Vat*, Cambridge: Cambridge University Press, pp. 155–173.

Heller, M. 1999. Relevant Alternatives and Closure. *Australasian Journal of Philosophy*, 77: 196–208.

Kallestrup, J. 2012. *Semantic Externalism*. Routledge.

Lewis, D. 1996. Elusive Knowledge. *Australasian Journal of Philosophy*, 74: 549–567.

McKinsey, M. 1991. Anti-individualism and Privileged Access. *Analysis*, 51: 9–16.

–, 2018. Skepticism and Content Externalism, *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2018/entries/skepticism-content-externalism/

Nuccetelli, S. 2003. Knowing That One Knows What One Is Talking About. In Nuccetelli, S. (ed.), *New Essays on Semantic Externalism and Self-Knowledge*, Cambridge Mass: The MIT Press, pp. 169–184.

Nozick, R. 1981. *Philosophical Explanations*. Harvard University Press.

Putnam, H. 1973. The Meaning of 'Meaning'. *The Journal of Philosophy*, 70: 699–711.

–, 1981. *Reason, Truth and History*. Cambridge University Press.

–, 1992. Replies. *Philosophical Topics: The Philosophy of Hilary Putnam* 20 (1): 347–408.

Thorpe, J. R. 2019. Semantic Self-Knowledge and the Vat Argument. *Philosophical* Studies, 176: 2289–2306.

Tymoczko, Th. 1989. In Defense of Putnam's Brains. *Philosophical Studies*, 57: 281–297.

Warfield, T. A. 1998. A Priori Knowledge of the World: Knowing the World by Knowing Our Minds. *Philosophical Studies*, 92: 127–147.

Wright, C. 1992. On Putnam's Proof that We Are Not Brains in a Vat. *Proceedings of the Aristotelian Society*, 92: 67–94.