SENSITIVITY, SAFETY, AND KNOWLEDGE FROM VIRTUAL REALITY

Alexandru DRAGOMIR, Mihai RUSU

ABSTRACT: The aim of this note is to analyze four externalist conditions on knowledge about the real world based on beliefs formed in VR. We will discuss Wheeler's sensitivity conditions for VR-based knowledge and propose a case wherein his favored condition, Virtual Sensitivity+, fails. Furthermore, we will advance two safety conditions and argue that while they pass our test case, they do not pass all of Wheeler's tests. We will conclude that none of the four conditions on VR-based knowledge about the real world is adequate, motivating a *prima facie* skepticism regarding the possibility of externalist conditions on VR-based knowledge.

KEYWORDS: virtual reality, externalism, reliabilism, sensitivity, safety

1. Wheeler's discussion of sensitivity conditions for VR-based knowledge about the real word

In his (2020) paper, Billy Wheeler tackled the problem of whether it is possible to acquire knowledge about the real world as a result of experiencing and interacting with virtual environments, i.e., computer-generated virtual objects and events. What motivated his inquiry is the fact that an influential externalist account of knowledge, i.e., Nozick's (1981) truth-tracking theory of knowledge, seems to exclude this intuitive possibility (Wheeler 2020, 369). According to the truth-tracking theory of knowledge, a necessary condition for knowing that P is that of *sensitivity*:

(Sensitivity) If P were false, S would not believe that P.

Or, in a formal manner, using '~' for negation, '$\Box\rightarrow$' as a sign denoting counterfactual conditionals, and 'Bs(P)' to denote 'agent S believes that P':

$$\sim P \Box\rightarrow \sim Bs(P)$$

Wheeler (2020, 369) notes that if S forms the belief that P is true in the real world (which we will formalize with 'Pr') based on forming the belief that P is true in VR (to be formalized as 'Pv'), then it is intuitive that in a nearby possible world in which Pr is false, S still believes that Pr is true, given that the VR makes S experience a virtual world in which Pv is true. Now, since (Sensitivity) does not

hold, it follows that we cannot acquire knowledge about the real world based on the beliefs we form in VR. After rejecting (Sensitivity), Wheeler considers a variation on it, inspired by McBain's (2017) take on knowledge in VR:

> (Virtual Sensitivity) If P were false in VR, then S would not believe that P is true in VR:
>
> $$\sim Pv \ \Box\!\!\rightarrow \sim Bs(Pv)$$

The core justification for using (Virtual Sensitivity) in a truth-tracking account of knowledge about the real world based on beliefs formed about VR is that the subject acquires knowledge about the real world only if the belief formed about how things are in VR is sensitive to how things are in VR. However, Wheeler shows that (Virtual Sensitivity) fails to provide a satisfactory standard of knowledge, as it permits (a) knowledge in cases that intuitively do not qualify as such, while (b) excluding knowledge in situations where it is plausible that agents would possess it.

Regarding (a), (Virtual Sensitivity) does not exclude knowledge in cases like the following, where S forms a true belief about the real world as a result of luck:

> COMPUTER MALFUNCTION: A new education VR program has gone to market. In its current form it contains a falsehood about the real world. Whereas P is true in the real world, P is false in the virtual world of the education program. S buys the program and runs it on their computer. However, their computer has a malfunction that incorrectly reads not-P as P, and so when is implemented, creates a visual experience of P. On this basis S comes to believe that P is true, both in the virtual world and the real world. (Wheeler 2020, 383)

According to Wheeler, "[i]n this case S has a true belief, both about the virtual world and the real world, and their belief about the virtual world is sensitive to the truths of the virtual world." (2020, 384) So what Wheeler claims is that the agent believes that Pr as a result of believing Pv, both of them being true – but is (Virtual Sensitivity) satisfied? Intuitively, the closest world in which Pv is false is the world before the malfunction, where the program is wrong by design. In this world, no malfunction occurring, ~Pv is read by the hardware as ~Pv. The agent forms the belief that ~Pv, so the agent will *not* believe that Pv. Consequently, (Virtual Sensitivity) is satisfied. Conceding that knowledge excludes luck, we have a case in which (Virtual Sensitivity) fails to conclude that the agent lacks knowledge about the real world.

Turning to (b), Wheeler notes that (Virtual Sensitivity) excludes knowledge in a situation where intuitively the agent acquires it:

> SPECTRUM INVERSION: A virtual reality program has been designed that deviates systematically from the real world. Every 5 minutes once per hour the virtual world inverts the colors that are experienced by a user. A user S has

> experienced this world many times and this has caused their brain to compensate for the inverted periods. During an inverted period, the virtual world displays a blue stop sign, however S comes to believe that the stop sign in the virtual world is red. On this basis S comes to form the belief that stop signs in the real world are also red. (Wheeler 2020, 384-5)

What happens in this case is that in the closest possible world in which Pv is false ('stop signs are not red in VR'), i.e., in the actual world, during the five minutes of color inversion, the agent forms the belief Pv is true ('stop signs are red in VR'). Thus, SPECTRUM INVERSION fails (Virtual Sensitivity), so counterintuitively implying that the agent does *not* acquire knowledge that stop signs are red.

After rejecting the McBain-inspired condition of (Virtual Sensitivity), Wheeler proposes a variation that passes the tests above, i.e., it denies knowledge about the real world in COMPUTER MALFUNCTION and grants it in SPECTRUM INVERSION:

> (Virtual Sensitivity+) If P were false about the real world, then S would not believe that P were true about the virtual world:
>
> $$\sim Pr \ \Box\!\!\rightarrow \sim Bs(Pv)$$

Wheeler argues that (Virtual Sensitivity+) excludes knowledge in COMPUTER MALFUNCTION. As Wheeler (2020, 387) notes, the agent believes that Pr and Pv, believing Pr as a result of believing Pv, but the belief formed about the VR world (i.e., the belief that Pv) is not sensitive to the real world (i.e., to Pr). Recall that Pr is actually true according to the description of the case, and suppose that it is false in the closest possible world. Then, in that world, the VR program - by design - is right in presenting Pv as false. However, the malfunctioning computer reads ~Pv as Pv, so the agent will form the belief that Pv is true. Putting things together, what happens is that in the closest world where Pr is false, the agent still forms the belief that Pv is true. Which contradicts the (Virtual Sensitivity+) condition, and, as such, excludes knowledge in COMPUTER MALFUNCTION.

Regarding SPECTRUM INVERSION, Wheeler (2020, 387) notes that in the closest worlds where Pr ('stop signs are red in the real world') is false (where the stop signs are, say, blue in the real world) the VR program typically presents virtual stop signs as blue, but every hour for five minutes it inverts the colors so that virtual stop signs appear red. S's brain compensates by inverting colors back, so S forms the belief that stop signs are blue, thus finally not believing that stop signs are red. Consequently, SPECTRUM INVERSION satisfies (Virtual Sensitivity+).

In the following we will argue that (Virtual Sensitivity+) is not a satisfactory condition for knowledge about the real world based on beliefs formed in VR. The case we propose is inspired by Sosa's (1999) CHUTE CASE, therefore subsequently

we will investigate similar safety conditions for VR-based knowledge of the real world, and find them defective. To wit, (Virtual Sensitivity+) is too strict, not allowing knowledge in a CHUTE CASE-like scenario, whereas what we will call (Virtual Safety) fails both of Wheeler's test cases. An improved version of (Virtual Safety), modeled on Wheeler's (Virtual Sensitivity+), which we call (Virtual Safety+), will still fail in one case, that is, it will not exclude knowledge in COMPUTER MALFUNCTION. This appears to motivate a skeptical conclusion regarding the possibility of providing a satisfactory externalist standard for knowledge about the real world based on beliefs formed in VR.

2. A discussion of safety conditions for VR-based knowledge about the real world

Sosa (1999) introduced the counterfactual condition of safety as an alternative to sensitivity. Both had an essential role in addressing the problem of luckily true beliefs, but sensitivity proved to be an excessive requirement, excluding knowledge in cases where the epistemic agent should intuitively be regarded as possessing it:

> [CHUTE CASE] On my way to the elevator I release a trash bag down the chute from my high rise condo. Presumably I know my bag will soon be in the basement. But what if, having been released, it still (incredibly) were not to arrive there? That presumably would be because it had been snagged somehow in the chute on the way down (an incredibly rare occurrence), or some such happenstance. But none such could affect my predictive belief as I release it, so I would still predict that the bag would soon arrive in the basement. My belief seems not to be sensitive, therefore, but constitutes knowledge anyhow, and can correctly be said to do so. (Sosa 1999, 145-6)

In Sosa's CHUTE CASE, the epistemic agent has good inductive reason to believe that the garbage bag is in the basement. Simply put, each time the bag was thrown, it fell down the chute and reached the basement. This is a good reason for us to assert that the agent knows that the garbage bag will reach the basement. However, sensitivity fails in the CHUTE CASE because, in the closest possible case in which the bag does not end up in the basement, the agent still holds the belief that it did arrive there. Here follows Sosa's safety condition:

> (Safety) S would not have believed that P without it being true[1], or
>
> If S were to believe P, P would be true:
>
> $$Bs(P) \mathrel{\Box\!\!\rightarrow} P$$

[1] This is Sosa's (1999, 146) reading of the safety condition.

Note that (Safety) holds true of the CHUTE CASE, since the closest possible world in which the agent believes that the garbage bag has arrived, i.e., the actual world, is a world in which the bag did indeed end up in the basement.

We will now describe a similar case regarding real-world knowledge based on beliefs formed in VR. In such a case, (Virtual Sensitivity+) appears to deliver the wrong result:

> [VIRTUAL EIFFEL TOWER] Using educational software based on a virtual world will likely help one acquire a good deal of knowledge about the real world. Say Mary uses a VR education tool aimed at teaching students about the history and geography of France. One of the classes includes a virtual tour of Paris and it teaches a bit about the history of the Eiffel Tower, presenting facts such as: it was inaugurated in 1889, having a height of 312 meters in 1889, but after antennas were added it reached 330 meters. As a result of her interaction with the VR world, Mary believes that the Eiffel Tower is 330 meters tall in VR. Since she is confident that the information being taught by the VR program is accurate, she also forms the belief that the Eiffel Tower is 330 meters tall in the real world. However, given variations in temperature and the plausible possibility of new antenna installations, it could have easily been 331 meters tall. In this situation, the agent would still believe that the Eiffel Tower is 330 meters tall in VR (correctly) and the real world (wrongly).

What we have is a case in which, in the actual world, Mary believes that the Eiffel Tower is 330 meters tall in both the virtual and real world. However, in the closest possible world in which the tower is slightly taller, she keeps believing that it is 330 meters tall in the real world. In other words, should the tower have been different in height in the real world, she would have still believed that it is 330 meters tall in VR. Consequently, Wheeler's (Virtual Sensitivity+) fails to obtain in this situation, although it is intuitive that Mary knows that the tower is 330 meters tall - in both the real and the virtual world. Let us note in passing that the previous condition that Wheeler explores, (Virtual Sensitivity), delivers the intuitively correct answer in VIRTUAL EIFFEL TOWER, since the closest possible world in which Pv is false is different from the actual world, and in such a world where Pv is false, one would not normally form the belief that Pv is true, all other things being equal.

One assumption of Wheeler's is that (Virtual Sensitivity+) meets the most plausible externalist requirements on knowledge based on beliefs formed in VR. However, the failure of (Virtual Sensitivity+) in the VIRTUAL EIFFEL TOWER case motivates us to consider an alternative safety condition. Let us start with a variant of Safety tailored to Wheeler's McBain-inspired variant of Sensitivity:

> (Virtual Safety) S would not have believed that P is true in VR without it being true

in VR:

$$Bs(Pv) \ \Box\!\!\rightarrow Pv$$

Regarding the VIRTUAL EIFFEL TOWER case, (Virtual Safety) delivers the right result: the closest possible world in which Mary believes that the tower is 330 meters tall in VR is the actual world, wherein the tower is exactly that height in the VR world. Returning to Wheeler's basic cases, we ask similarly if (Virtual Safety) does the job of excluding knowledge in COMPUTER MALFUNCTION. The answer is that it does not: the closest world in which the agent believes that Pv is the actual world, and in the actual world, as a result of the malfunction, Pv is true. Moreover, (Virtual Safety) excludes knowledge in SPECTRUM INVERSION: the closest possible world in which S believes that Pv (S believes that stop signs are red in VR) is the actual world, where stop signs are not red in the VR environment as a result of malfunctioning. So (Virtual Safety) fares no better than (Virtual Sensitivity).

Now let us entertain a safety-like condition on knowledge grounded in VR-based beliefs, tailored on Wheeler's (Virtual Sensitivity+):

(Virtual Safety+) S would not have believed that P is true in VR without it being true in the real world:

$$Bs(Pv) \ \Box\!\!\rightarrow Pr$$

First, let us note that (Virtual Safety+) gets right the test of SPECTRUM INVERSION, allowing that the agent has acquired knowledge that stop signs are red: the closest world in which the agent believes that Pv (that stop signs are red in VR) is the actual world, wherein real-world stop signs are red.

Now, let us see whether (Virtual Safety+) delivers the right result in the VIRTUAL EIFFEL TOWER case. The closest world in which Mary believes that the Eiffel Tower is 330 meters in VR is the actual world, wherein the Tower is indeed 330 meters tall. What this entails is that (Virtual Safety+) does not exclude knowledge of the fact the tower is 330 meters tall in the real world, based on the belief that the virtual version of the tower is 330 meters tall, and this is what we should plausibly expect.

However, we should note that (Virtual Safety+) does not exclude knowledge in the COMPUTER MALFUNCTION case. The closest world in which the agent forms the belief that Pv is the actual world (after the malfunction, ~P is read as P), wherein Pr also holds true.

3. Conclusion

Starting from Wheeler's (2020) paper, we have analyzed four externalist conditions on VR-based knowledge of the real world, and concluded that none offers the right

results in all three test cases. In the following table we summarize the results, using '+' to indicate that a condition delivers the intuitive result, and '-' to indicate that it does not:

	COMPUTER MALFUNCTION	SPECTRUM INVERSION	VIRTUAL EIFFEL TOWER
Virtual Sensitivity	-	-	+
Virtual Sensitivity +	+	+	-
Virtual Safety	-	-	+
Virtual Safety +	-	+	+

One possible solution would be to combine the best candidates, i.e., (Virtual Sensitivity+) and (Virtual Safety+). Under a conjunctive approach, knowledge requires the satisfaction of both conditions. If at least one of these conditions is not satisfied, the subject lacks knowledge. Since, as argued above, (Virtual Sensitivity+) is not satisfied in the VIRTUAL EIFFEL TOWER case, the conjunctive approach entails the counterintuitive conclusion that the subject lacks knowledge in this case. A disjunctive approach posits that knowledge is attained if at least one of the conditions is satisfied. The trouble is that the disjunctive approach leads to the counterintuitive conclusion that the subject has knowledge in the COMPUTER MALFUNCTION case. An alternative solution is to segregate the application of conditions, that is, to claim that (Virtual Sensitivity+) should be applied to some types of cases, whereas (Virtual Safety+) should be applied to other cases. However, this appears to be an *ad hoc* solution, without any substantial philosophical motivation behind it. This discussion lends credence to a skeptical view of the possibility of formulating a universal externalist standard for knowledge of the real world based on beliefs formed in VR. Although this appears to be an important and challenging topic, defending an overarching skeptical conclusion would be beyond the scope of this discussion note, but could become the subject of further investigation.[2]

References

McBain, J. 2017. "Epistemic Lives and Knowing in Virtual Worlds". In *Experience Machines: The Philosophy of Virtual Worlds*, edited by Mark Silcox, 155–68. Lanham: Rowman & Littlefield International.

[2] We would like to thank Marian Călborean, Andrei Mărășoiu, Emilian Mihailov, Elena Popa, and Corina Stavilă for their comments on previous versions of this paper. Alexandru Dragomir was supported by a grant of the Research Institute of the University of Bucharest (ICUB, nr. 10504/25.10.2024) "The effects of LLM interaction on TOM in digital and virtual environments".

Nozick, R. 1981. *Philosophical Explanations*. Cambridge, MA: Harvard University Press.

Sosa, E. 1999. "How to Defeat Opposition to Moore". *Noûs* 33 (s13): 141–53. https://doi.org/10.1111/0029-4624.33.s13.7.

Wheeler, B. 2020. "Truth Tracking and Knowledge from Virtual Reality". *Logos & Episteme. An International Journal of Epistemology* 11 (3): 369–88. https://doi.org/10.5840/logos-episteme202011327.