Appendix B: Experimental-Philosophy-Style Surveys on AI's First Premise

As reported in Section 8 of Chapter 2, when, with Joshua Knobe's help, I asked on a couple of experimental-philosophy-style surveys (726 participants in total) whether people thought they knew they were not BIVs, only 41% chose 'I don't know that I'm not a BIV', while 59% chose 'I know that I'm not a BIV'. These numbers represent the results of two surveys, taken over three days.[1]

In the first survey, 215 participants were recruited using Amazon's Mechanical Turk and given this description:

> Let's use "BIV" (for brain in a vat) to mean a brain that has no body, but is being kept alive in a vat, and is hooked up to a super-advanced computer that sees to it that all aspects of a normal brain's interactions with its body and with the world around it are perfectly simulated. So everything seems to the BIV exactly as it would if it had a body and were experiencing an external world.

They were then asked:

> We would first like to ask the following question, which is designed to see if you understood what a BIV is.
>
> If a BIV, as just described, were having the experience of eating a blueberry, how would that seem to the BIV?

And almost all (94%) of respondents chose 'It would seem to the BIV exactly as if it had a body and was eating a blueberry', with only 6% instead choosing the other option provided, 'It would seem to the BIV just a little bit different from how eating

[1] I should note that my surveys drew quite a few more male than female respondents, which makes me wonder about that and also about potential other ways that the pool might not be representative of the general population. Of the 720 respondents who gave their gender (6 didn't), 63% were male, while only 37% were female. I should note that there seemed to be a significant difference between the genders, with females being more inclined than males to respond that they do know that they're not BIVs: 68% of women said that they knew, while 54% of men said that they knew. This difference was statistically significant, $\chi 2$ (1, N = 720) = 13.4, $p < .001$. However, caution is advisable before concluding that there is a significant gender difference on this and other philosophical questions; on this, see (Seyedsayamdost 2014).

a blueberry would seem to a normally embodied brain'. On the main question, respondents were told 'We are interested in whether you can know that you are not a BIV of the type just described,' and then asked which of the two options they thought correctly described them, and 58% chose 'I know that I'm not a BIV', and 42% chose 'I don't know that I'm not a BIV'. Thus, people's tendency to say that they knew they were not a BIV was significantly greater than what would be expected by chance alone, χ2 (1, N = 215) = 5.1, p = .02.

In a second survey, 511 participants were given this slightly different description:

> Let's use "BIV" (for brain in a vat) to mean a brain that has no body, but is being kept alive in a vat, and is hooked up to a super-advanced computer, that, taking into account the motor output of the BIV, gives the BIV appropriate sensory input. Because all aspects of a normal brain's interactions with its body and with the world around it are perfectly simulated in a BIV, everything seems to a BIV exactly as it would if it had a body and were experiencing an external world.

They were asked the same initial question as in the first survey, and this time 91% answered that things would seem to the BIV exactly as they would if it had a body and were eating a blueberry. On the main question, results were quite similar to the first survey: this time 60% chose 'I know that I'm not a BIV' and 40% chose 'I don't know that I'm not a BIV'. Again, this is significantly greater than chance, χ2 (1, N = 511) = 20.8, p < .001.

As I remarked in Chapter 2, these are very different from the much more skeptic-friendly results I had earlier obtained by the quite different means of taking a show of hands among students in an introductory philosophy class. (See Chapter 2, Section 8 for discussion of these differences.) This difference in results is rendered even more remarkable by the confidence that their answer was right that was reported by those who answered 'I know that I'm not a BIV' on the later x-phi-style surveys. After answering the main question, I asked respondents, 'How confident are you that your answer to the previous question is correct?', on a scale of 1 to 7, with 7 labeled as 'most confident' and 1 as 'least confident.' Figure 1 shows the distribution of confidence levels for participants who answered that they did know that they were not BIVs.

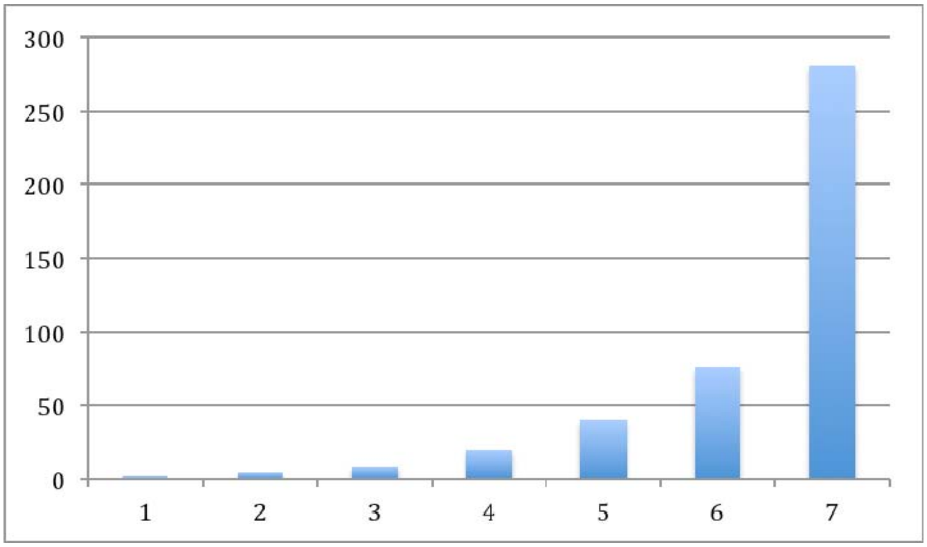

Figure 1. Number of participants at each confidence level, among those who said that they did know they were not BIVs, collapsing across the two studies.

I find it arresting that on a question where there is a somewhat close 59%-41% split in answers, so many of those who gave the majority answer would be so confident that they are right.[2] (Recall, though, that each person is answering whether they take themselves to know that they're not a BIV. This opens the possibility—however slight—that most everybody is right: Maybe many really do know this of themselves, while many others don't!) And this confidence of so many that they are right to think they know makes it seem, at least to me, even more remarkable that in the different setting of my asking students at the start of a philosophy course what they think, so very many would say that they do *not* know.

I have long been bothered by the confidence philosophers often project on our answers to questions we're in no position to know the answer to. (See Appendix C for related discussion.) These results tempt me to the thought that it's people generally who tend to be overconfident when they deal with philosophical questions—and that this afflicts philosophers more than others just because we spend more of our time on these matters.

[2] The participants who said that they did not know (295 in total) that they were not BIVs were on the whole less confident that their answer was correct. The distribution was as follows: 36% (106) chose 7; 21% (63) chose 6; 18% (53) chose 5; 16% (48) chose 4; 4% (11) chose 3; 2% (6) chose 2; and 1% (4) chose 1. Thus, there is a significant effect such that those who say that they do know show higher confidence levels, $t(719) = 8.0$, $p < .001$.