Inside the Matrix: Imperfections in Introspection

Jonathan Riedel

Follow this and additional works at: http://commons.colgate.edu/car

Part of the Philosophy Commons

Recommended Citation
Available at: http://commons.colgate.edu/car/vol4/iss1/11
Inside the Matrix: Imperfections in Introspection

Jonathan Riedel, 2009

This paper was written for a class entitled Philosophy of Psychology. The main objective of the class is to tackle the question of what to do when neuroscientific evidence conflicts with what ordinary people report of their experience. What happens when the personal explanation is at odds with the subpersonal? One such manifestation of the conflict, dubbed “the interface dilemma,” is introspection: what one reports about their feelings and beliefs, barring a motivation to lie, seems incorrigible. No one can tell me I am wrong when I say I am sad because the only access to that knowledge is through introspection, my private and personal experience.

When I mention terms like “autonomist” and “neuro-computationalist,” I am referring to the different schools of thought about this issue. Autonomists, on one end of the spectrum, believe that the accuracy of first-person reports is inviolable. For them, when I say I am in love and my brain patterns don't match up to what everyone else's do when they are in love, 'love' has to be redefined in the neuroscientist's dictionary, not in people's ordinary usage; people trump science. On the other end of the spectrum are the neuro-computationalists (or the eliminativists), who believe that progressive scientific discoveries will eventually tell us what we need to know about the brain, such that currently ambiguous matters about thoughts, beliefs, and qualitative experience will have a clear definition, much the way that the “folk” notion of the four humors was eventually put to rest by the discovery of microbes as the primary cause of illness. More moderate views exist, but in this context the battle between extreme views is the major concern.

Philosophy of Psychology, and in particular the question of whether introspection is reliable, is very important for how we live our daily lives. Should we trust a scientific report more than a person telling to our faces the details of his or her experience? A serious challenge to introspection, something we take to be so basic to our understanding of ourselves and the world, is definitely a matter worth discussing, and I hope my advocacy of Nisbett and Wilson will allow us to hold on to that faculty while taking into consideration the undoubtedly auspicious progress of science.

What if The Matrix were real? It is certainly provocative to entertain the thought that we are living in a world that we think to be real but is actually not. Despite its intrigue, the sane of us generally regard the film’s premise as false. There is no way that we can be wrong about something so simple as that which is right in front of our noses at all times, and it is even more menacing to think that even our own mental states are subject to inaccuracy – right? Sadly,
scientific studies lead to theories presented by Gopnik, Gordon, and Nisbett and Wilson that we are in a kind of prison of the mind, where introspection is a flawed means of accessing the truth about mental states. Of these particular theorists, however, Nisbett and Wilson present the most clear-cut challenge to introspection without leaving us in absolute skepticism like Gopnik and Gordon do. While all of the above would agree that what we assume to be introspection is imperfect, Nisbett and Wilson argue not to the effect that introspection itself is wrong, but that our conception of it is often mistaken. Instead, the explanation of behavior resides in a priori causal relationships. So, while Gopnik and Gordon take blows at the foundations of epistemology, the latter pair presents a serious challenge to introspection without destroying all talk about the interface dilemma and knowledge in general.

The discussion of behavior explanations begins with simple psychological experiments. For example, the answer to “how did you recall your mother’s maiden name?” is usually along the lines of “I don’t know, it just came to me.” Such an inarticulate answer for this and for similar questions suggests that perceptual and memorial processes have a nebulous underlying structure; perhaps our inability to explain the “higher-order processes” for recalling such facts means that introspection is not a good means of identifying our mental states. This anti-introspectivist view seems problematic for a few reasons, according to Nisbett and Wilson: (1) Inarticulate explanations about perception and memory do not guarantee that all higher-order processing is chaos; (2) Even though questions about recall maybe answered poorly, people generally know how to articulately answer other kinds of questions like “why they behaved as they did in some social situation or why they like or dislike an object or another person”; (3) Since introspection is utterly and completely wrong, people can never be correct about their mental states, a highly troubling assertion if certain states are only accessible by this means.

The basis of Nisbett and Wilson’s overall argument is the challenge of these seeming problems. The writers claim that: (1’) Some connections between stimuli and effects which are clearly affecting an experiment go unnoticed by subjects, meaning that “the accuracy of subjective reports is so poor as to suggest that any introspective access that may exist is not sufficient to produce generally correct or reliable reports”\(^1\); (2’) Not just questions about recalling maiden names are subject to confusion – even explanations of preference or acting in social situations may not be given as a result of introspection but what is wrongly assumed to be introspection; (3’) People can be correct, but usually they are correct accidentally. In the most powerful blow to knowledge via introspection, (2’), Nisbett and Wilson


\(^2\) Ibid., 233.
note that people who claim to introspect on reasons for an effect are often in fact using “implicit, a priori theories about the causal connection between stimulus and response.” In (3’), they assert that people are correct when what would be true as determined by introspection is also true as determined by using the aforementioned sub-personal knowledge of cause and effect relationships. Such accuracy usually occurs when the given information is limited and explicit.

Why, now, should we believe that we are not good at introspecting? Like Gopnik, Nisbett and Wilson show substantial evidence from a plethora of experiments that (1’) is true. Nisbett’s own data on shock subjects, who performed differently depending on whether they had taken a placebo, suggests that deeper cognitive processes affect subjects’ behavior – in this case, tolerance of pain – while their own explanations of their behavior report otherwise. The subjects claim that they were not thinking about the pill at all, but their behavior suggests that at an unconscious level they were using knowledge of having taken a pill. If the implications of such experiments are correct, then it is clear that subjects “cannot report on the existence of influential stimuli. It therefore would be quite impossible for them to describe accurately the role played by these stimuli in influencing their responses.”

Their second and third premises deal with statements of the sort, “I came to a stop because the light started to change” and “I played a trump because I had no cards in the suit that was led.” Because the explanation (the part of the sentence after the word “because”) is recalled from memory based on the questions “why did you come to stop?” and “why did you play a trump?”, it feels like a definite result of introspective access. One looks for the answer in some history of mental states and gives an explanation based on what his or her state was that seemed to produce the effect. Nisbett and Wilson, however, assert that the explanations are not actually given via introspection. Rather, our knowledge of pre-existing causal relationships override whatever we understand to be introspection: “the culture or a subculture may have explicit rules stating the relationship between a particular stimulus and a particular response.” What is perceived as the cause of one’s behavior is actually already set. While it is true that I come to a stop because the light started to change, I know that I do this because the relationship between my stopping a car and my seeing a yellow light is already established, and I base my explanations off this relationship.

This position, in my view, is convincing because it provides relief from the injustice of a five year-old’s favorite game: asking perpetual “why” questions. The existence of a priori

---

3 Ibid.
4 See p. 253, When Will We Be Correct In Our Verbal Reports?
5 I am using the term “unconscious” literally, i.e. the subjects themselves did not recognize the possibility that they were behaving based on having taken the pill.
6 Ibid., 240.
7 Ibid., 248.
8 Ibid.
causal theories means a rest from *ad infinitum* interrogation of the form:

- “Why did you stop?”
- “Because the light started to change.”
- “Why did the light’s changing make you stop?”
- “Because the law says I should stop.”
- “Why does the law’s saying so motivate you to stop?”

Etc.

In the direct introspective way of evaluating behavior, there is never really an end to why one acts the way one acts. If (2') and (3') are true, one knows the way one acts is based on causal relationships, and the explanation stops there. Although it may feel like introspection because the information is retrieved internally, it is not. What actually counts as introspection, then, is an independent issue. Nisbett and Wilson’s argument allows for introspection to still be valid, but that our distinction between using introspection and using a priori theories is wrong. Such a claim is ultimately and auspiciously benign.

Furthermore, the argument from Nisbett and Wilson lacks nothing in seriousness, despite its acceptable implications. Like Gopnik and Gordon, they still challenge an individual’s ability to introspect, but the result is not threatening to the idea of knowing the reasons for one’s behavior. In the other authors’ views, it seems that anything we say about explanations for actions is wrong if it comes about internally. Gopnik extrapolates her theory to the broadest degree, questioning whether correct introspection for anyone is even comprehensible. If young children think they are right but are actually so clearly wrong, how do adults know that they are not wrong? In Gopnik’s world, perhaps we are living in the matrix. Gordon’s idea of an egocentric shift is implausible for reasons too expansive to be discussed here.

Nisbett and Wilson, however, do not say that all introspection is wrong, simply that what we think to be introspective access is in fact something else. The division between what “feels like introspection” and “introspection itself” is existent but its boundaries are nebulous. Since they assume that theory-laden belief production is far more prevalent than actual introspection, they still assert that introspective access is not as easy and accurate as one might think. When we try to explain our behavior and get inarticulate results in some cases – “it just felt like it”, “it just came to me” – or articulate results in others – “I played a trump card because...” – we realize that we still can be right but are often not. The conditions under which we are correct “accidentally” are too elaborate to discuss here (see footnote 4).

As a serious challenge to the existence of direct access to mental states, Nisbett and Wilson’s theory has dramatic implications for the interface problem. Autonomy theorists seem to come out on top; if sub-personal explanations (e.g. those discovered via introspection) are flawed, then the autonomists lose nothing in attributing full autonomy to person-level explanations. Furthermore, Nisbett and Wilson’s idea that what we take to be introspection is actually a subtle recall of a priori causal theories supports

---

9 (3’) covers this possibility.
autonomy theory because then person-level relations are, at the core, the means to accessing mental states. The idea that personal and sub-personal explanations have no questions in common to answer is perfectly commensurable with Nisbett and Wilson’s theory. Predictably, the neuro-computationalist loses one battle even though the personal-sub-personal distinction is conspicuously sharpened. That certain kind of sub-personal explanation which we rely on so heavily to be accurate is indeed flawed. However, the neuro-computationalist also wins a battle: if people are unaware of their reliance on a priori causal theories in almost 100% of cases – i.e. people wrongly believe they are introspecting – then commonsense psychology is in trouble.

In the full-blown war, autonomy theory can be said to win. Even Nisbett and Wilson themselves speak out against the stubborn reclusiveness of neuro-computational eliminativism. Even if it feels like introspection is “nothing more than judgments of plausibility”, they argue, there are other factors that make it unfair to discredit all introspective reports and ignore all arguments based on them.\(^\text{10}\) Autonomy theorists would probably support the use of introspection, as long as it avoided affecting person-level characteristics, while the eliminativists would ignore one side altogether.

Independent from its effects on the interface dilemma, the challenge of Nisbett and Wilson that introspection is not exactly what we thought it was is nothing to turn a nose up at. In fact, the reason the challenge is so compelling is because it can exist without questioning all access to knowledge. If explanations of our behavior rely on pre-existing cause and effect relationships, so what? We would certainly like to believe we are right when we internally access a mental state, but if we are not correct, there does not seem to be any harm done. The damage may be done to our egos, but at least there is a way to make sure we are not part of the matrix.

\(^{10}\) Ibid., 255.