Observing a Superposition*

Paul Skokowski
St. Edmund Hall, Oxford University
paul.skokowski@seh.ox.ac.uk

3 March, 2021

Abstract

The bare theory is a no-collapse version of quantum mechanics which predicts certain puzzling results for the introspective beliefs of human observers of superpositions. The bare theory can be interpreted to claim that an observer can form false beliefs about the outcome of an experiment which produces a superpositional result. It is argued that, when careful consideration is given to the observer's belief states and their evolution, the observer does not end up with the beliefs claimed. This result leads to questions about whether there can be any allure for no-collapse theories as austere as the bare theory.

Introduction

The bare theory is an intriguingly simple Everettian interpretation of quantum mechanics which is explored by Jeff Barrett in *The Foundations of Quantum Mechanics* (Barrett 2020) and *The Quantum Mechanics of Minds and Worlds* (Barrett 1999), and by David Albert in *Quantum Mechanics and Experience*. (Albert 1992) The idea behind the bare theory is straightforward:

*Forthcoming in *Synthese*. Please cite the published version.

"The bare theory is just the standard collapse formulation of quantum mechanics but without the collapse dynamics." (Barrett 2020, p. 145) By removing the collapse postulate from the von Neumann-Dirac version of quantum mechanics, "...the linear dynamical laws are nonetheless the complete laws of the evolution of the entire world." (Albert 1992, p. 123) Quantum mechanical states will thus evolve deterministically according to the dynamics of the Schrödinger equation.

However, when a human observer gets involved – and in particular when that person observes a superpositional state – the bare theory appears to lead to baffling results. For example, Barrett maintains that "...the bare theory seeks to explain why one might *falsely believe* that one had determinate appearances of the sort predicted by the standard theory,"(Barrett 1999, p. 110, emphasis Barrett's), and "If the bare theory were true...an observer would typically believe that she had an ordinary determinate experience when there would in fact be no such experience that she believed she had."[1](Barrett 1999, p. 112) As Albert puts it, such an observer "will be radically deceived even about *what her own occurrent mental state is*"(Albert 1992, p. 118), and "Whatever belief [the observer] *does* end up with ... is necessarily going to be a false belief." (Ibid., p. 127) These are claims about belief. In particular, these claims concern the content of the observer's introspective state. A careful analysis of the nature of belief within the quantum mechanical formalism – either the bare theory or the von Neumann-Dirac version – will need to address the complicated intentional aspects of belief, the contents of belief, and the neural components of belief, including its causal roles and vehicles/eigenstates. This paper will begin by providing such an analysis of belief and introspection in the context of an observer of spin $1/2$ outcomes. Next it will apply this analysis to superpositional outcomes in the context of the bare theory. It will be shown that the observer has no false belief in the form claimed. It will then be shown that the quantum mechanical property of linearity cannot produce such a result given

[1]I have replaced "he" with "she" in this quote.

the fine-grained nature of belief contents, and the time evolution of the belief eigenstates. This result will be reinforced with an example of a simple spin measurement system. The example will show how a common output obtained from a superposition lacks the properties required to count as a misrepresentation about pointer position. This is because such common outputs lack the fine-grainedness and resolution to represent any pointer positions in the first place. These results will lead us to question whether there can be any allure for no-collapse theories as austere as the bare theory.

1. The bare theory, introspection, and superposition

In their discussions of the bare theory, both Albert and Barrett consider the example of a human observing a superpositional state which has resulted from a Stern-Gerlach measurement. I will follow Barrett's notation, as it uses more standard coordinates.[2] The example considers an observer "M" who measures the x-spin of a spin $1/2$ system S.[3] Call this the $M{+}S$ system. Before interacting with the measuring device, the system S is in an eigenstate of z-spin, and the observer M is in an eigenstate of being ready to measure the x-spin of the system S. Post-measurement, a superpositional state of the observer M and the spin $1/2$ system S, will result from the linear dynamics of the bare theory. The resulting superposition is given by Barrett's equation (4.1), written here as equation (1):[4]

$$|\psi\rangle = \frac{1}{\sqrt{2}}\Big(|x\text{-spin up}\rangle_M \, |\uparrow_x\rangle_S + |x\text{-spin down}\rangle_M \, |\downarrow_x\rangle_S \Big), \tag{1}$$

which Barrett shortens to:

$$|\psi\rangle = \frac{1}{\sqrt{2}}\Big(|\uparrow\rangle_M \, |\uparrow_x\rangle_S + |\downarrow\rangle_M \, |\downarrow_x\rangle_S \Big).$$

[2]Albert uses 'hardness' to denote spin along the x-axis, and 'color' to denote spin along the z-axis. Barrett uses traditional Cartesian coordinates x, y, and z for the spin directions. I have replaced Albert's terminology of "hard" with "up", and "soft" with "down", to be consistent with the spin terminology being used in this paper.

[3]For the purposes of this discussion, we shall consider the spin $1/2$ particles to be electrons.

[4]Barrett's notation combines observer and measuring apparatus for M, assuming perfect correlation between the two. (Barrett, 1999) See also the discussion in section 2 below.

As mentioned above, the bare theory is most puzzling when it is applied to mental states like beliefs, and so it's worth underscoring that the observer M is a *human* observer. In a separate discussion, Barrett describes an automaton that records spin measurements, where "This model requires a close correspondence between physical memory configurations and mental states. . . ." (Barrett 1999, p. 95) But presumably automatons (and models of automatons) do not themselves exhibit mental states such as beliefs and experiences, and so it is the human mental states that arise in the observation of a superposition which need explaining, as both Barrett and Albert themselves recognize. For example, in the first sentence of the section entitled *The account of experience*, Barrett asks "So just how far can the bare theory go in explaining our experience?", where "our" refers to human experience (1999, p. 110), and later "This is what it *feels like* to be" the observer, and that "she will *believe*" that the pointer indicates a determinate result. (Barrett 2020, p. 147, Barrett's emphasis) Albert similarly asks "what it would feel like to be the experimenter" (Albert 1992, p. 116) and in particular asks the experimenter "whether or not you have any particular belief" (Albert 1992, p. 118) about the outcome. Albert goes on to stipulate that M's perceptual eigenstates are belief states *within M's brain* that track the pointer, and thus have content, and he explicitly labels these internal states as *belief* eigenstates, for example: $|\text{believes } e \textit{ up}\rangle_M$.[5] These are not simple descriptions of external behavior, but instead are descriptions of the internal representational states of a human observer *M* - internal states which feel a certain way to her, which have representational contents such as pointer positions, and which have causal consequences. In other words, these are descriptions of human *mental* states, and in particular, belief states.[6] (Kim 2010; Dretske 1988; Dretske 1995) We will focus then, on the details of human mental states such as beliefs when a superposition occurs as a

[5] Albert 1992, p. 112. See also the discussion in Section 2 and footnote 14, below. Here I have replaced "black" with "up", and "*h*" with "M".

[6] As Albert and Barrett's remarks reveal, what is most fascinating about M observing a superposition is what she ends up *believing* about the experiment. And behaviorist approaches to understanding her mental states (including belief) will invariably end up short, as they leave out the vehicles, contents (e.g., pointer readings) and causal roles of these mental states.

result of applying the bare theory.[7]

Consider now a human observer of an experiment where the outcome is a superposition as in equation (1). Barrett's question for the observer M in the state is "Did you get a determinate result of either x-spin up or x-spin down?" (Barrett 1999, p. 98) Albert also asks the subject if she has "...any particular definite belief...about the value of the [spin] of this electron." (Albert 1992, p. 118) Note that asking a person to report in this way on the content of a belief they hold requires introspective access to that belief. One must introspect in order to access the existent belief, and thus report the content of that belief. Barrett agrees, saying that M "...would believe that she knows what the result is." (Barrett 1999, p. 98) M's *belief* about what she *knows* is an introspective belief, and in this instance, M is being asked to introspect her perceptual belief/knowledge about the result of the experiment. This introspection is a belief about a belief.[8] And so it is M's introspection which is misrepresenting what she is perceiving.

Consider a simple, and non-superpositional, case where M perceives a red apple. We would say in this situation that M has the occurrent perceptual belief that the apple is red. Such an occurrent belief would involve M's visual cortex. (Zeki 1993; Lee et al. 1998; Seymour et al. 2016) Were we to now ask M about the color of the apple in front of her, M would presumably report "The apple is red." But now let us ask M whether she has a determinate result for her perceptual belief about the color of the apple before her. M can rightly respond to this query with "Are you asking me what color the apple is?" to which, following Albert and Barrett's prescription, we would answer, "No. We are asking a different question. The question is, do you now have a determinate perceptual belief about the color of the apple before you?

[7] Albert and Barrett's key claim about the bare theory concerns false introspective beliefs attributed to observer *M*. Neither author attempts to clarify any distinctions between mental state terms such as 'experience' and 'belief'. Since their key claim concerns only belief, we will likewise focus on the properties of beliefs when evaluating their claims about the bare theory, as properties of experience will not weigh on their conclusions.

[8] As knowledge states are typically taken to be a form of true belief then M's belief about her knowledge state is a belief about a belief; hence, an introspective belief. From here on for consistency we will refer to M's perceptual beliefs of the experimental outcomes as beliefs rather than knowledge.

Specifically, we are asking you to verify that you have a determinate belief about the color of the apple - by introspecting your belief about the color of the apple." In this instance, M will employ an introspective belief – a belief about a belief – because she will need to inspect her occurrent beliefs to establish that she indeed has a belief about a red apple before her. As an introspective belief, this will be a belief with another belief as a content; and specifically, M's occurrent perceptual belief will be the content of M's introspective belief.

We should note that in all the cases Albert and Barrett describe, the initial occurrent beliefs M forms about the device reading are occurrent *perceptual* beliefs about the device, and that M is then tasked to introspect those perceptual beliefs. Thus the mental states in question are perceptual beliefs, and introspections of perceptual beliefs. It is a hallmark of perceptual mental states that their contents are fine-grained.[9] These fine-grained contents ensure that M's mental state – be it a perceptual belief, an introspection of a perceptual state, an experience, etc. – about (or of) a red apple will always be distinct from her perceptual belief, introspection of a perceptual belief, or experience about (or of) a green apple. (Tye 1995; Tye 2009; Dretske 1988; Dretske 1995; Frege 1892; Perry 1977) The same lesson applies when M observes the detector in her experiment: the fine-grained contents of M's mental states ensure that M's belief, introspection, or sensation that the electron has spin up will be distinct from her belief, introspection, or sensation that the electron has spin down.[10] This fine-grainedness is fundamental: a belief about redness has an intentional content that differs from the intentional content of a belief about greenness. The intentional content of a belief that the needle registers "$+1/2$" is distinct from the intentional content of a belief that the needle registers "$-1/2$". Beliefs with different intentional contents will always be distinct from each other. Hence, any type of mental

[9] We will see that the fine-grainedness of M's mental states is enforced in three ways: by results from neuroscience and through an observational principle applied by Albert and Barrett (both in Section 2), and by the property of transparency of mental states (Section 6).

[10] Where the content in these cases includes, say, the position of a pointer on the measuring device; for example pointing to one of '$+$' or '$-$'.

state with a spin-up content will differ from any mental state with a spin-down content.[11] And importantly, it is the contents of the states that give us the license to call such states mental representations to begin with. (Brentano, 1874; Dretske 1988; Tye 1995) These contents help us distinguish one state from another, and in causal theories of mental content, mental states are individuated through their differences in content, vehicle, and causal role. (Dretske 1988; Kim 2010)

Perception and introspection are also mental states that involve different physical regions of the brain. Visual perception, as already pointed out, involves visual cortex. Introspection involves pre-frontal cortex, according to imaging studies. (Fleming et al., 2010) When introspection is being utilized to report an occurrent perceptual belief, the two beliefs – introspective and perceptual – have distinct neural vehicles located in different regions of the brain. Each neural vehicle is made up of neurons exhibiting their own set of action potentials during the introspective/perceptual belief episode. When asking M to report about one of her occurrent perceptual beliefs, her answer will depend, as we have seen, on the introspective state which is representing that occurrent perceptual belief. In addition, the intentional content of M's introspective state will have a fine-grainedness that tracks the fine-grainedness of the occurrent perceptual belief it represents.[12] (Dretske 1995; Moore 1903; Tye 2009) This means simply that when asked, if M is visually perceiving the needle pointing to x-spin up, then M will introspect this perception of the needle pointing to x-spin up. If alternatively, M is visually perceiving the needle pointing to x-spin down, then M will introspect this perception of the needle pointing to x-spin down.

Returning briefly to the baffling results of the bare theory, the claim is that M "...would

[11] Note that this difference in content holds whether the content is the position of a pointer towards '$+$' or '$-$' or whether the content is actual color content 'red' or 'green.' The intentional content will be fine-grained in either case.

[12] If the representing state did not have this fine-grainedness, then M would not be capable of answering queries about the content of the perceptual state in question.

typically believe that she had an ordinary determinate experience when there would in fact be no such experience...", and "Whatever belief M *does* end up with, when (1) obtains, is necessarily going to be a false belief." It is this claim of M's having a particular false belief that deserves addressing, since M's having this false belief is what makes the theory so baffling in the first place. Second, since M has an introspective belief about an occurrent perceptual belief of a spin result, then it is M's *introspective belief* about this occurrent perceptual belief which is false. We will revisit these results below, after first coming to understand the details of M's observing x-spin measurements in both non-superpositional, and superpositional, cases.

2. The non-superpositional case

Let us begin by considering a non-superpositional case where M perceives, and introspects, an x-spin up result for system S. Representing M's introspection of her belief that the electron is x-spin up will require two eigenstates: one corresponding to her introspective state involving pre-frontal cortex, and another corresponding to her perceptual belief involving visual cortex. The state of a non-superpositional system corresponding to M's introspection of her perceptual belief, together with an observed x-spin up system S, can be written:

$$|\text{Introspect} \uparrow\rangle_{M-PF} |\uparrow\rangle_{M-VC} |\uparrow_x\rangle_S . \tag{2}$$

These states represent not just M's occurrent perceptual belief about the spin result, $|\uparrow\rangle_{M-VC}$, but also her introspection of that perceptual belief, $|\text{Introspect} \uparrow\rangle_{M-PF}$. Here the subscript "$M-PF$" stands for M's introspective state involving neurons in Pre-Frontal cortex and the subscript "$M-VC$" stands for M's perceptual belief state involving neurons in Visual Cortex.

Similarly, the state of a system corresponding to M's introspection of her perception of a non-superpositional x-spin down result will be:

$$|\text{Introspect} \downarrow\rangle_{M-PF} |\downarrow\rangle_{M-VC} |\downarrow_x\rangle_S . \tag{3}$$

It is important to note that M's perceptual and introspective beliefs are stipulated to be perfectly accurate: she "... is a perfect observer of the measurement result indicated by the measuring device and the measuring device is perfect in correlating the position of the pointer that represents its result with the x-spin of S." (Barrett 2020, p. 106) This means that "... whenever M looks at a pointer that's pointing to "up," she eventually comes to believe that the pointer is pointing to "up"; and that whenever M looks a pointer that's pointing to "down," she eventually comes to believe that the pointer is pointing to "down" (and so on, in whatever direction the pointer may be pointing)."[13] (Albert 1992, p. 77) Let's refer to M's perfectly accurate observations of Stern-Gerlach results as the *Accuracy Principle*.

The Accuracy Principle means that the contents of M's beliefs correspond exactly with the spin of an electron as measured and displayed by the device M is observing. For example, if a prepared x-spin up electron is passed through a detector aligned to measure spin in the x-direction, then the electron will emerge in the x-spin up state, causing the arrow on the device to point to spin up, and M will perceive the arrow on the device pointing to spin up, with the resultant perceptual belief of the measurement as x-spin up. The resulting M+S non-superpositional system is:

$$|\uparrow\rangle_{M-VC} |\uparrow_x\rangle_S .$$

For a prepared x-spin down electron, the resulting M+S non-superpositional system is:

$$|\downarrow\rangle_{M-VC} |\downarrow_x\rangle_S .$$

This means there will be specific neurons in visual cortex that will fire when the perceived arrow points to x-spin up, and that a separate and distinct set of neurons in visual cortex will fire (with different action potentials) when the perceived arrow points to x-spin down. These two sets of neurons will differ in location in visual cortex, and will differ in their action potentials. This

[13] Barrett refers to observer M as "him" which I have changed to "her" in the quote. Albert calls M "Martha". I have replaced Albert's terminology of "hard" with "up", and "soft" with "down" in this paper.

accords with experimental results from neuroscience, where single neurons and small groups of closely clustered neurons in visual cortex have preferences to respond to directional and shape properties of a perceived object, with different neurons responding to distinct directions and shapes. For example, directionality bias in single and closely clustered neurons in visual cortex has been demonstrated with experiments on rhesus monkeys. This bias is fine-grained, with specific small clusters of neurons being biased for particular, preferred, directions. (Salzman et al., 1990; Salzman et al., 1992) In experiments on macaque monkeys, single neurons were found to give selective responses to specific shapes and directions. (Tanaka et al. 2003; Tanifuki et al. 2001; Dehaene 2009) Again, an eigenstate representation such as $|\uparrow\rangle_{M-VC}$ captures that this is a physical state in M's brain[14] – hence the subscript "$M-VC$" denoting a specific state in Visual Cortex – and that this is the state of perceiving x-spin up.

And, when M is *introspects* this perceptual result of x-spin up, we have:

$$|\text{Introspect}\uparrow\rangle_{M-PF}\,|\uparrow\rangle_{M-VC}\,|\uparrow_x\rangle_S\,.$$

Here, the physical state representation $|\text{Introspect}\uparrow\rangle_{M-PF}$ captures that this is a physical state in M's brain, hence the subscript "$M-PF$" denoting a specific state in M's Pre-Frontal Cortex which is the introspective state of the perception of x-spin up.

Similar results hold *mutatis mutandis* for an x-spin down electron sent through the detector:

$$|\text{Introspect}\downarrow\rangle_{M-PF}\,|\downarrow\rangle_{M-VC}\,|\downarrow_x\rangle_S\,.$$

We are now in a position to consider the time evolution of the "ready" state of a non-superpositional system before the measurement to the state after the measurement. Before a measurement of a

[14] Albert stipulates that M's states $|\uparrow\rangle_{M-VC}$ and $|\downarrow\rangle_{M-VC}$ are physical perceptual belief eigenstates in M's brain that track the pointer of the measuring device: "...|"up"$\rangle_M$ is that physical state of M's brain in which she believes that the pointer is pointing to the word "up" on the dial... and |"down"$\rangle_M$ is that physical state of M's brain in which she believes that the pointer is pointing to the word "down" on the dial."(Albert; 78) Here I have replaced Albert's observer "o" with "M", and have replaced "hard" with "up" and "soft" with "down".

prepared x-spin up electron, the state of the M+S system is:

$$|\text{PF ready}\rangle_{M-PF}\,|\text{VC ready}\rangle_{M-VC}\,|\uparrow_x\rangle_S\,, \tag{4}$$

where the first state is the 'ready' state of M's Pre-Frontal cortex, the second state is the 'ready' state of M's Visual Cortex, and the third state is the (system S) electron prepared in the x-spin up direction. Since M is a perfect observer of electron spin outcomes, her state evolves after the measurement into the familiar result (2):

$$|\text{Introspect}\uparrow\rangle_{M-PF}\,|\uparrow\rangle_{M-VC}\,|\uparrow_x\rangle_S\,.$$

The case is similar *mutatis mutandis* for the time evolution of states for a prepared x-spin down electron. Before the measurement, the state of the system is:

$$|\text{PF ready}\rangle_{M-PF}\,|\text{VC ready}\rangle_{M-VC}\,|\downarrow_x\rangle_S\,. \tag{5}$$

And, since M is a perfect observer, her state evolves into the familiar result (3):

$$|\text{Introspect}\downarrow\rangle_{M-PF}\,|\downarrow\rangle_{M-VC}\,|\downarrow_x\rangle_S\,.$$

These non-superpositional results are the same for both the von Neumann-Dirac formulation and the bare theory. With only a single possible outcome, the collapse postulate gives the same result as the deterministic Schrödinger equation. In addition, any measurements of observables for the individual eigenstates within either of these cases will give the same results in both formulations. So, M perceives x-spin up when an x-spin up electron has been passed through the device, and x-spin down when an x-spin down electron has been passed through.

These outcomes accord precisely with the fine-grainedness of mental states. We would expect, given what we know about visual cortex, that a perception of a pointer pointing to "up" would have a distinct neural vehicle exemplified in visual cortex from the neural vehicle exemplified in a perception of a pointer pointing to "down". But we also expect, given the nature

of belief, that the perceptual content of spin-up will always be distinct from the perceptual content of spin-down. And because M is a perfect observer of spin results, then the contents of M's perceptions will always be the results displayed on the measuring device, in whatever form that device is set up to display: "+", or "−"; "↑", or "↓"; "up" or "down", etc. So when it comes to belief states in non-superpositional cases, both their neural vehicles/eigenstates and their contents are fine-grained, and track/agree with each other with respect to the spin of the system, and this result is guaranteed by the Accuracy Principle.

Similarly, if M introspects the an x-spin-up result, then that means that M is perceiving x-spin up, and is introspecting this result. Since M is a perfect observer, and because introspective states are belief eigenstates, then when M introspects a perceptual state of a spin outcome, the introspective states and their contents will track the spin results. This outcome also accords with the fine-grainedness of mental states. So an introspection of a perception of spin-up, for example a pointer pointing to "up", will have a neural vehicle exemplified in pre-frontal cortex distinct from the neural vehicle exemplified in pre-frontal cortex of an introspection of a perception of spin-down. And given the nature of belief, the introspection of a perceptual content of spin-up will always be distinct from the introspection of a perceptual content of spin-down. So, as with other belief states, both the neural vehicles and the contents of introspections of perceptions of spin outcomes are fine-grained, and agree with each other with respect to the spin of the system, and this result is guaranteed by the Accuracy Principle.

3. Observing a superposition in the bare theory

Now consider the time evolution of M's pre-frontal and visual cortex when the system S is in a superposition. Before measurement, the system S is in an eigenstate of z-spin, and the observer

M is in an eigenstate of being ready to perceive, and introspect, the result of the measurement:

$$|\text{PF ready}\rangle_{M-PF}\,|\text{VC ready}\rangle_{M-VC}\,\Big[\frac{1}{\sqrt{2}}(|\uparrow_x\rangle_S + |\downarrow_x\rangle_S)\Big],$$

where the first state is the 'ready' state of M's Pre-Frontal cortex, the second state is the 'ready' state of M's Visual Cortex, and the state in brackets is the superposition of x-spin up and x-spin down (an electron initially in an eigenstate of z-spin expanded in the x-spin basis) for the electron that will be passed through the detector.

Reformulating terms:

$$\frac{1}{\sqrt{2}}\Big[\,|\text{PF ready}\rangle_{M-PF}\,|\text{VC ready}\rangle_{M-VC}\,(|\uparrow_x\rangle_S + |\downarrow_x\rangle_S)\Big]$$

$$= \frac{1}{\sqrt{2}}\Big[\,|\text{PF ready}\rangle_{M-PF}\,|\text{VC ready}\rangle_{M-VC}\,|\uparrow_x\rangle_S + |\text{PF ready}\rangle_{M-PF}\,|\text{VC ready}\rangle_{M-VC}\,|\downarrow_x\rangle_S\,\Big]$$

We recognize that the first component in brackets is the state (4) and the second term is the state (5) which were introduced earlier. So both these components will evolve according to the Schrödinger equation as above, and the Accuracy Principle will ensure that the physical and content properties track each other within each component. The result after the measurement will be a superposition which we will call $|\Psi\rangle$, of the form:

$$|\Psi\rangle = \frac{1}{\sqrt{2}}(|\text{Introspect}\uparrow\rangle_{M-PF}\,|\uparrow\rangle_{M-VC}\,|\uparrow_x\rangle_S + |\text{Introspect}\downarrow\rangle_{M-PF}\,|\downarrow\rangle_{M-VC}\,|\downarrow_x\rangle_S).$$

Both components of the superposition have introspective and perceptual eigenstates. As discussed earlier, these eigenstates each represent distinct neural vehicles and contents. Any measurement of these eigenstates would yield distinct eigenvalues one from the other. Since each belief eigenstate in the superposition is distinct from the others, and since the eigenstates of belief are orthonormal, the superposition cannot correspond to any determinate state of belief.

Now we are in a position to note two important things. First, equation (1) does not accurately describe the state of the system where M introspects her belief about the measurement outcome.

Instead, the state of the system will be given by state $|\Psi\rangle$, which contains a superposition of M's introspective eigenstates involving pre-frontal cortex and perceptual belief eigenstates involving visual cortex. Second, finding a common, measurable eigenvalue for the belief eigenstates in the superposition $|\Psi\rangle$ that M is in post-measurement, appears to be a formidable task, given that each eigenstate represents distinct physical, and content, properties.

Both Albert and Barrett propose a solution to this difficulty. The solution is to get M to answer "Yes" to a specific question about her mental state. Albert asks for M to answer whether she has "... any definite belief ... about the value of the [spin] of this electron." (Albert 1992, p. 118) Barrett formulates the question for M as, "Did you get some determinate result to your x-spin measurement, either x-spin up or x-spin down?" (Barrett 1997, p. 98) And so, Barrett continues, "M would report that she got a determinate x-spin result when she did not determinately get up and did not determinately get down." (Ibid)

By linearity, any answer to this question by M would need to be the same given in a non-superpositional case. We have seen that each component of $|\Psi\rangle$ represents a non-superpositional case: the first component in $|\Psi\rangle$ is state (2) and the second is state (3):

$$|\text{Introspect} \uparrow\rangle_{M-PF} |\uparrow\rangle_{M-VC} |\uparrow_x\rangle_S$$

$$|\text{Introspect} \downarrow\rangle_{M-PF} |\downarrow\rangle_{M-VC} |\downarrow_x\rangle_S$$

When in either state (2) or (3), M will need to introspect her perceptual belief about the result of the experiment in order to evaluate the question of whether she has perceived x-spin up or x-spin down. And an evaluation of the question requires access to the particular occurrent perceptual belief in each case. Since the occurrent perceptual beliefs in either case are distinct from one another in content ("up" in one component and "down" in the other), then the contents of the introspections in either case will also be distinct from one another. In the first case, M's introspective evaluation of the perceptual belief that the spin is up would reveal the answer to

be "up", whereas in the second case, M's introspective evaluation of the perceptual belief that the spin is down, would reveal the answer to be "down". Let us designate the distinct contents of these two evaluative introspective belief states as "UP $\vee$ DOWN?: UP" and "UP $\vee$ DOWN?: DOWN". M requires intentional contents with values like these in order to answer the question correctly in either case. And both these cases occur in different components of the superposition $|\Psi\rangle$. Given these considerations, let us say that upon formulating these evaluative introspective belief eigenstates, M can now answer the question as posed: "Did you get some determinate result to your x-spin measurement, either x-spin up or x-spin down?"

But note that there is a peculiarity about answering such a question. Spoken answers to questions are formulated in a different part of the brain: Broca's area. And Broca's area is an area of the brain that is tasked with linguistic output – *not* with introspection. These linguistic outputs include unconscious grammatical processing, together with sending the signals required to form the mouth and tongue in a particular configuration, exhaling breath in a certain manner, opening and closing the nasal passages, and so forth.(Pinker 1994, 1997) Further, linguistic processing is unconscious (Pinker 1994; Tononi 2012), whereas introspection is conscious (Dretske 1995; Moore 1903). Finally, introspection is tasked with producing beliefs about beliefs, and has the function to indicate the contents of the beliefs being introspected (Dretske, 1995), whereas Broca's area is not: it is tasked with producing linguistic outputs. (Pinker 1994, 1997) Since linguistic output states have different neural vehicles, types of contents (unconscious vs conscious) and functional roles from introspective states, they require different eigenstate representations, which will be considered in the next section.

The outputs of intentional mental states should not be confused with mental states themselves. Consider the simple example of drinking a beer. I believe there's a six-pack in the fridge and I desire a beer. These mental states cause me to open the fridge door, grab a bottle, twist off the cap, and take a drink. The belief and desire are mental states with intentional contents,

but the reaching, twisting, and drinking are outputs which are caused by these representational states. These outputs are not themselves representational states: states with intentional contents from a function to represent properties in the environment, and executive capacities to cause action.(Dretske 1988; Papineau 1987) They are instead *outputs* of representational states: causal consequences of mental states which *do* have the executive capacities and intentional contents.

The questions that have been posed, therefore, are designed to detect a common measurable output of introspective states in the M+S system, and not directly measure the introspective and perceptual eigenstates themselves. Yet what is at issue is the content of M's own introspective and perceptual beliefs, for the claim is that M has a false introspective belief. We should note that M is not being asked in these questions to introspect what spin result she perceived. Indeed, Albert explicitly commands M not to state what she believes is the actual spin of the electron: "Don't tell me whether you believe the electron to be spin up or you believe it to be spin down, but tell me merely whether or not one of those two is the case."(Albert, 118) The reason for this prescription is clear: The eigenvalue for asking M "Do you introspect you are perceiving spin up?" will be different for both components of the superposition, as it will if we decide instead to ask M "Do you introspect you are perceiving spin down?", and so there will be no common eigenvalue for the superposition $|\Psi\rangle$ if either of these questions is asked. So a different question – a disjunctive one – must be asked ("whether or not one of those two is the case"). But that means that, rather than measuring the contents of M's introspective states directly, we are instead being asked to measure a common output of those introspections. The focus has been shifted from introspection itself – and so a question about M's beliefs – to a common *output* of introspective beliefs.

4. An example

The problem of shifting the focus in this way can be illustrated by an example. The example shows how a single output can also be observed from a superposition involving an x-spin measuring device. We will see that the eigenstate which produces the single result lacks the fine-grained contents of other eigenstates in the superposition – the pointer states that represent fine-grained spin results – even though it is an output of these very states. This will lead us to reconsider how to interpret M's report when including output states in her Broca's area.

It is worth emphasizing that when an electron-spin measuring device ends up pointing to '$+$', or pointing to '$-$', these pointer states are representational states with a fine-grained content: they are *about* something, and crucially, they are *about* whether the measured electron is spin up or spin down. That's what makes them representational in a way appropriate for accurately measuring the spin of a particular electron.

Suppose that the pointer on our electron-spin measuring device starts in a "ready" position pointing straight up as depicted in figure 1. When an electron is passed through the device, the pointer either rotates to the left for a spin-down result, or to the right for a spin-up result.

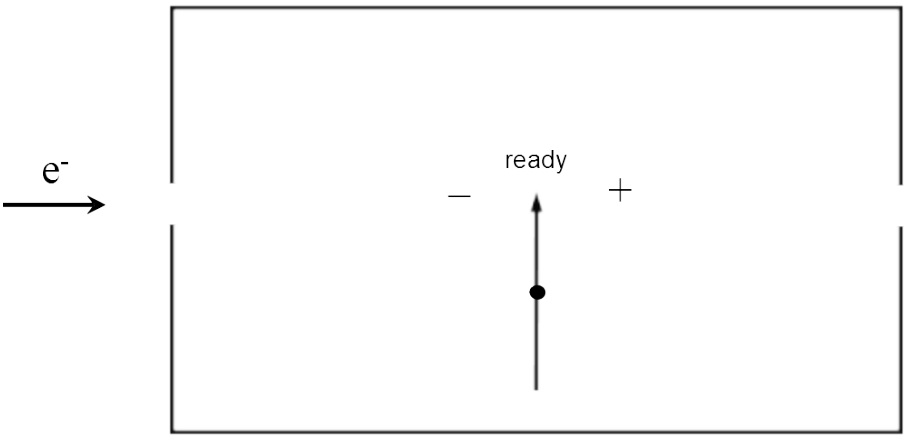

Figure 1: A simple spin-measuring device

To make the analogy with our human observer clear, call the spin measuring device "M". Use the subscript "M-P" to designate the pointer eigenstate. Before the measurement we have:

$$|\text{ready}\rangle_{M-P}\,[\frac{1}{\sqrt{2}}(|\uparrow_x\rangle_S + |\downarrow_x\rangle_S)],$$

where the first state is the 'ready' state of the pointer M-P, and the state in brackets is the superposition of the x-spin up direction and x-spin down direction (that is, the electron in system S initially in an eigenstate of z-spin expanded in the x-spin basis) for the prepared electron.

The result after the measurement will be a superpositional state we will call $|\Phi\rangle$:

$$|\Phi\rangle = \frac{1}{\sqrt{2}}(|\uparrow\rangle_{M-P}\,|\uparrow_x\rangle_S + |\downarrow\rangle_{M-P}\,|\downarrow_x\rangle_S)$$

Now let us attach a circuit to the measuring device M, as shown in figure 2.

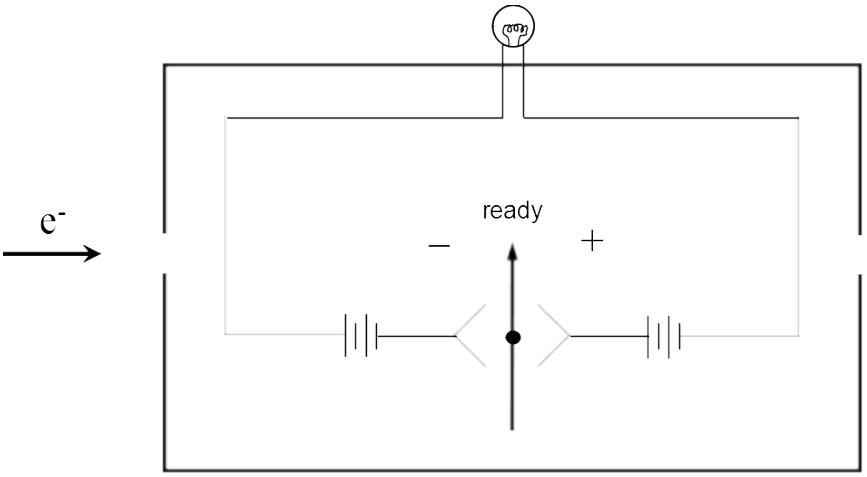

Figure 2: Measuring device M with circuit, ready to measure a spin

When the needle comes to rest, pointing to either spin direction, it closes the circuit shown, causing a flow of current which turns on a lightbulb, as shown in figures 3 and 4 below.

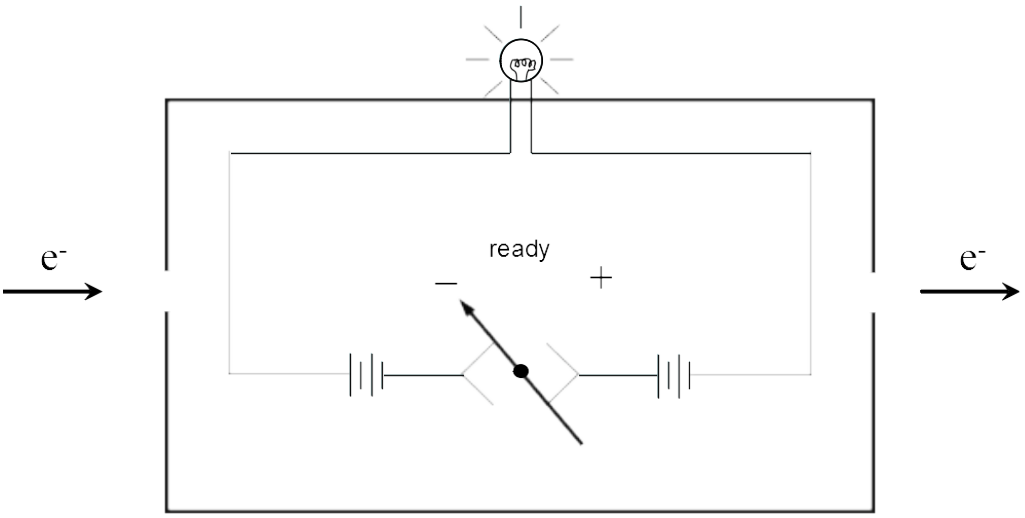

Figure 3: A measurement of x-spin down.

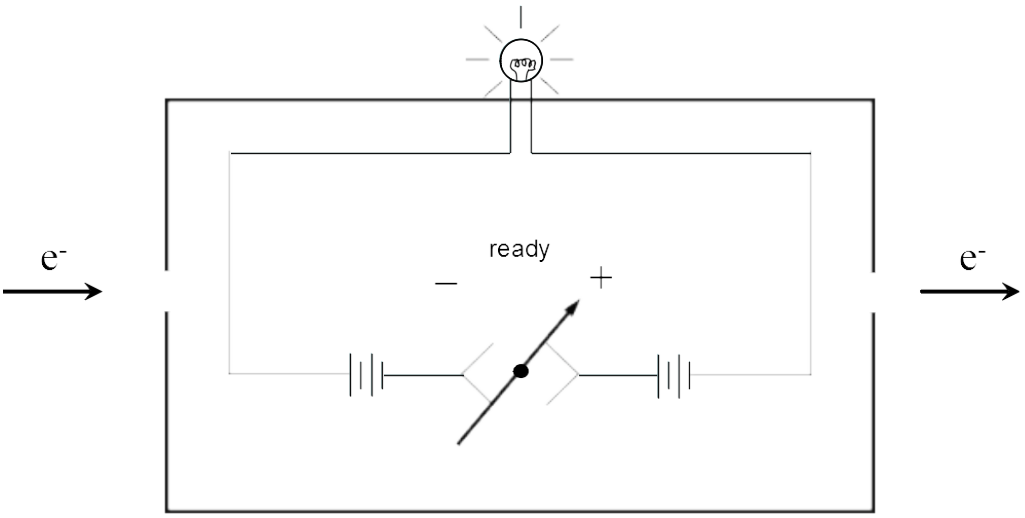

Figure 4: A measurement of x-spin up.

Denote the lightbulb eigenstate with the subscript "M-L". Before measurement we have:

$$|\text{ready}\rangle_{M-L}\,|\text{ready}\rangle_{M-P}\,[\frac{1}{\sqrt{2}}(|\uparrow_x\rangle_S + |\downarrow_x\rangle_S)],$$

where $|\text{ready}\rangle_{M-L}$ and $|\text{ready}\rangle_{M-P}$ are the ready eigenstates of the measuring device M. The result after the measurement will have the lightbulb M-L now being on, represented by the

lightbulb eigenstate $|\text{“On”}\rangle_{M-L}$, and will yield the superpositional state we will call $|\Phi'\rangle$:

$$|\Phi'\rangle = \frac{1}{\sqrt{2}}\Big(|\text{“On”}\rangle_{M-L} |\uparrow\rangle_{M-P} |\uparrow_x\rangle_S + |\text{“On”}\rangle_{M-L} |\downarrow\rangle_{M-P} |\downarrow_x\rangle_S \Big).$$

Call the light intensity when the lightbulb goes on from the completed circuit, λ. Then we can see, by linearity, that a measurement of this lightbulb output by an operator O is an observable property of the superpositional state as well as an observable of each component of the superposition. So, by putting an electron in an eigenstate of z-spin through an x-spin detector, we can measure this observable property λ:

$$\begin{aligned} O|\Phi'\rangle &= \frac{1}{\sqrt{2}} O\Big(|\text{“On”}\rangle_{M-L} |\uparrow\rangle_{M-P} |\uparrow_x\rangle_S + |\text{“On”}\rangle_{M-L} |\downarrow\rangle_{M-P} |\downarrow_x\rangle_S \Big) \\ &= \frac{1}{\sqrt{2}} \Big(O|\text{“On”}\rangle_{M-L} |\uparrow\rangle_{M-P} |\uparrow_x\rangle_S + O|\text{“On”}\rangle_{M-L} |\downarrow\rangle_{M-P} |\downarrow_x\rangle_S \Big) \\ &= \frac{1}{\sqrt{2}} \Big(\lambda|\text{“On”}\rangle_{M-L} |\uparrow\rangle_{M-P} |\uparrow_x\rangle_S + \lambda|\text{“On”}\rangle_{M-L} |\downarrow\rangle_{M-P} |\downarrow_x\rangle_S \Big) \\ &= \lambda|\Phi'\rangle . \end{aligned}$$

Thus it is possible to observe a single value, λ, from the superposition $|\Phi'\rangle$. But notice that measuring a lightbulb output like this does not mean that the device M is falsely representing the spin outcome in some way. It just means that the lightbulb emits a light flash regardless of whether there has been a collapse to one component of the superposition $|\Phi'\rangle$ or the other (as in the von Neumann/Dirac formulation), or whether, if Everett is correct, this observable will be measurable *even in a superposition*, as a consequence of linearity. As Albert points out:

> . . . it follows from the linearity of the operators that represent observables of quantum-mechanical systems. . . that if any observable O of any quantum-mechanical system S has some particular determinate value in the State $|A\rangle_S$, and if O also has that same determinate value in some other state $|B\rangle_S$, then O will necessarily *also* have precisely that same determinate value in any linear *superposition* of those two states. (Albert 1992, p. 117)

A measurement of lightbulb output λ not only does not, but *cannot*, represent any final position of the pointer, and hence the spin of the electron. If you are looking at the lightbulb for *that* fine-grained content, you are looking in the wrong place. The lightbulb can never represent this fine-grained content, as it does not have the resolution of the pointer. It has no "+" or "–" markings, no "↑" and "↓" markings, and no way of representing spin directions in the first place. As Dretske has put it, the lightbulb doesn't have the "function to indicate" such contents. (Dretske 1988; Dretske 1995) No measurement of the lightbulb can produce such a fine-grained output, just as no operator can operate on the lightbulb $|\text{“On”}\rangle_{M-L}$ eigenstate and produce "+" or "–" eigenvalues. The lightbulb is blind to these fine-grained contents, and simply cannot represent them. And if it can't represent these outputs, it can't mis-represent them either. (Dretske 1988; Dretske 1995) The lightbulb, in short, cannot falsely represent spin directions, as it has no capability to represent them in the first place.

5. Linearity and belief

The lightbulb/detector example is not a perfect analogue of the puzzling case presented by the observer M, because M has introspective states, whereas the lightbulb/detector system presumably does not; but we will see shortly that it is nevertheless instructive. To begin, recall that Albert and Barrett ask M to introspect her perceptual belief, and that the false belief she is claimed to end up with is an introspection, since M "...would believe that she knows what the result is." Again, M's *belief* about what she *knows* is an introspective belief. To see the difference between M's case and the lightbulb/detector, consider once more the case where M is in a *non*-superpositional state. In this situation, when M perceives a particular spin outcome, say "+", she will introspect that she is perceiving "+":

$$|\text{Introspect} \uparrow\rangle_{M-PF} |\uparrow\rangle_{M-VC} |\uparrow_x\rangle_S ,$$

and as we have seen, this is the result expected from the Accuracy Principle and the neuroscience of perception. But this result also occurs because of the transparency of mental states, which will be discussed here and in the following section. Briefly put, introspective states track and reveal the content of the occurrent perception which is being introspected. Introspective states have the function to indicate the contents of the states being introspected. (Dretske 1995) No additional features of the perceptual state are revealed in introspections of perceptual contents; the perception is *transparent*, yielding its perceptual content to introspection. (Moore 1903, Dretske 1995, Tye 1995, Tye 2000, Tye 2009) This is why, when M introspects her observation of spin up, "+", she introspects the content "+". This of course makes sense when we are asked, as M is, what the content of an occurrent perceptual belief is. The introspection provides the content for answering that question, and it does it by checking the content of the occurrent perceptual belief.

This situation does not occur for our simple lightbulb/detector, as we can see by considering this system in a non-superpositional state, such as when the detector detects the single spin outcome "+". This situation is depicted as:

$$|\text{“On”}\rangle_{M-L}\,|\uparrow\rangle_{M-P}\,|\uparrow_x\rangle_S$$

Suppose we now 'ask' the system to *introspect* the result of its perception of the spin of the electron. What could serve as the introspective state? The electron state $|\uparrow_x\rangle_S$ is what is *perceived*, so it will not do. The pointer state $|\uparrow\rangle_{M-P}$ plays the role of the occurrent perceptual state, as it displays the electron spin result.[15] This leaves us with the lightbulb state, $|\text{“On”}\rangle_{M-L}$. But the lightbulb being "On" cannot be a state of introspecting the pointer position "+", for several

[15]The pointer state cannot serve an an introspective state for another reason: introspective states are not identical with the states they introspect. As Dretske explains, when addressing the notion of 'introspective' states of instruments, "We can see the instrument's pointer positions. It cannot." (Dretske, 1995, p. 48) See his chapter 2, "Introspection," for why instruments and gauges can't introspect their own readings.(Ibid) We have already seen an example of this, for humans at least: introspective states are located in a different region of the brain from perceptual states.

reasons. First, as discussed in the previous section, the lightbulb cannot represent the content of the pointer, as it does not have the resolution of the pointer; it lacks the fine-grainedness of an introspective state tracking an occurrent perceptual state. It has no "+" or "–" markings, no "↑" and "↓" markings, and no way of representing spin directions whatsoever. Second, as just discussed, an introspective state will track and reveal only the intentional content of the occurrent perceptual state being introspected.[16] But here the contents of the two states are completely different: the pointer content is "+", whereas the lightbulb's content is "On". Additionally, if the pointer instead pointed to "–", the lightbulb's content wouldn't change at all - it would again have the content "On". These are not cases of introspection, since not only does the presumed introspective state of the lightbulb not track and reveal the content of the state being 'introspected,' but its content doesn't even change when the content of the state being introspected *does*! The reasons for these failures were given in the previous section: the lightbulb is blind to the fine-grained contents of the pointer state and does not have the function to indicate them, and so cannot represent *or* misrepresent them. The lightbulb state is instead an output of the pointer's states, and it altogether lacks the intentional properties required to be an introspective - or analogue-introspective - state.

And now we can recognize the same situation of an output state lacking the intentional properties of an introspection with the *human* observer M. Recall that M answers a question using Broca's area. When we consider the linguistic output state in Broca's area, then before the measurement the "ready" state of the $M-S$ system is:

$$|\text{B ready}\rangle_{M-B}\,|\text{PF ready}\rangle_{M-PF}\,|\text{VC ready}\rangle_{M-VC}\,[\frac{1}{\sqrt{2}}(|\uparrow_x\rangle_S+|\downarrow_x\rangle_S)], \tag{6}$$

where the first state is the 'ready' state of M's Broca's Area, and the rest of the 'ready' states are defined as before: M's pre-frontal cortex, her visual cortex, the 'ready' state of the measuring

[16] In "...bringing to bear your faculty of introspection...you are not aware of any *inner* object or thing. The only objects of which you are aware are the external ones making up the scene before your eyes." (Tye 2000, pp. 46-47) See also Dretske (1995), Moore (1905), and Tye (1995) and (2009).

device, and the final state is the superposition of up and down spin for the electron that is about to be passed through the detector. And as before, M's eigenstates are designated by subscripts '$M-B$', '$M-PF$' and '$M-VC$'.

After the measurement, this state evolves into a superposition of the form $|\Psi'\rangle$, which is different from the superposition $|\Psi\rangle$ we considered earlier, and now includes the output eigenstate $|\text{“Yes”}\rangle_{M-B}$ in M's Broca's area:

$$|\Psi'\rangle = \frac{1}{\sqrt{2}}\Big(|\text{“Yes”}\rangle_{M-B} \, |\text{Introspect} \uparrow\rangle_{M-PF} \, |\uparrow\rangle_{M-VC} \, |\uparrow_x\rangle_S$$
$$+ |\text{“Yes”}\rangle_{M-B} \, |\text{Introspect} \downarrow\rangle_{M-PF} \, |\downarrow\rangle_{M-VC} \, |\downarrow_x\rangle_S \Big).$$

Then we can see, by linearity, that a measurement of M's linguistic output will be an observable property of the superpositional state as well as of each component of the superposition. That is, when the superposition $|\Psi'\rangle$, which includes M's eigenstates, is asked whether M has some definite belief in the way prescribed earlier, where this question is represented by the operator O, then she will answer "Yes":

$$O \, |\Psi'\rangle = \text{“Yes”} \, |\Psi'\rangle \, .$$

And this result is by virtue of this operator O operating on M's state $|\text{“Yes”}\rangle_{M-B}$.

But this is like the detector example above. 'Measuring' an answer like this – an *output* of an introspective state – does not mean that M's *introspection* is false. It just means that M produces an output "Yes" regardless of whether she has collapsed to one component of the superposition or the other, or that, if Everett is correct, this answer will be measurable even in a superposition (by virtue of linearity). As was the case for the spin detector above, a measurement of an output "Yes" does not convey the content of *any* of M's introspective eigenstates or occurrent perceptual belief eigenstates about the spin of the electron; those states have very different contents, as we have seen. Indeed, M's belief eigenstates have not changed at all from superposition $|\Psi\rangle$ to superposition $|\Psi'\rangle$. This is explicitly spelled out in the state vector $|\Psi'\rangle$,

which contains M's four belief eigenstates: $|\text{Introspect} \uparrow\rangle_{M-PF}$, $|\uparrow\rangle_{M-VC}$, $|\text{Introspect} \downarrow\rangle_{M-PF}$, and $|\downarrow\rangle_{M-VC}$. Measurement of a non-belief eigenstate of M like $|\text{“Yes”}\rangle_{M-B}$ does not mean M's introspection eigenstates are misrepresentations, as these are separate eigenstates with their own associated operator (which commutes with the operator “O”) and eigenvalues. In addition, the contents of the belief eigenstates remain fine-grained with their contents determined (and unchanged) due to the Accuracy Principle. Recall from Section 3 that M's introspective evaluation of the question *requires* – due to the Accuracy Principle – that the content of any of M's introspective states will contain either an “up” or a “down” component. So when we consider the claim that M “would believe that she knows what the result is” (Barrett 1999, p. 98) based on this spoken output, and that this belief is “false” (Ibid.), we see that there is no basis for this claim, as M does not actually have the belief. That is, there is no single belief state that emerges from, or can be factored out of the superposition with the singular content spin-up or spin-down. Further, the spoken answer is not a belief state, and in particular it is not an introspective state as required by the bare theory, and so it has no intentional content. So M has no false introspection. There is instead a common linguistic output which occurs regardless of the quantum mechanical interpretation.

Bub, Clifton and Monton (1998) came to a similar conclusion regarding M's introspective states. They recognized that M's introspective states (what they call M's ‘reflecting’, or ‘believing that she believes’) would indeed differ in each component of the superposition, by tracking the perceptual belief's content in the same component:

> But if the bare theory is true, we can also ask what it will predict when M attempts to reflect upon what belief about e-spin she has. Since M would then get into a superposition of believing that she believes up and believing that she believes down... under the bare theory she will be unable to specify which of the two beliefs

she takes herself to hold.[17] (Bub et al. 1998, p. 42)

Though Bub et al. recognized this problem, they did not press the argument further, instead opting to criticize the bare theory on other grounds, including the inability of the bare theory to account for the ordinary beliefs that observers come to have about measurements.[18] This paper, in contrast, argues that Bub et al. should have continued the argument, for the two reasons given above. First, an utterance of "Yes" by M is an *output* of belief eigenstates, and so does not qualify as a belief; and second, due to the Accuracy Principle, no remaining *belief* eigenstate of M has a singular content that could possibly allow that state to serve as the required false belief.

In addition, it is important to recognize that within each component of the superposition $|\Psi'\rangle$ involving M, there are only two belief eigenstates: an introspective belief eigenstate and a perceptual belief eigenstate. These are M's *only* candidates for a false belief. However, due to the Accuracy Principle, none of these belief eigenstates could serve as instances of a false belief about tracked spin direction within that component. So M's belief eigenstates about pointer position considered individually – perceptual and introspective – are accurately tracking the spin direction *within that component*.

Further, none of M's *belief* eigenstates can be factored out of the superposition $|\Psi'\rangle$. If any could, we would have a candidate for a single belief state for M that might serve as the required false belief. Perhaps such a state would look something like $|\Psi'\rangle = |D\rangle\,|\Psi''\rangle$, where $|D\rangle$ is the factored-out eigenstate which is the false belief state, and $|\Psi''\rangle$ contains the superpositional 'residue', if you will, of $|\Psi'\rangle$. But none of the four existing belief eigenstates within the superposition $|\Psi'\rangle$ – namely, $|\text{Introspect}\uparrow\rangle_{M-PF}$, $|\uparrow\rangle_{M-VC}$, $|\text{Introspect}\downarrow\rangle_{M-PF}$, and $|\downarrow\rangle_{M-VC}$ – can be factored out of the superposition. So none of these eigenstates can serve as the single

[17] Bub refers to his subject as "Eve." "Eve" is replaced with "M" here.

[18] See also Barrett (1998) for more on Bub et al.'s arguments against the bare theory.

belief for M which could be considered to be the false belief $|D\rangle$. This means any false belief about the outcome for the superpositional case must be at the level of speech output; that is, at the level of M saying "Yes" about introspecting a definite belief about the spin of the electron. But speech output, as we have shown, is not an introspective state, and indeed, being an output rather than an intentional state, it is not a belief of any kind about the spin of the electron, and so it cannot be the claimed false belief.

Finally, we should also note that since knowledge is generally taken to be some form of justified true belief, then *M* cannot have any sort of *knowledge* of the outcome of the experiment, either. This potentially undermines a separate claim of Albert's that when observing a superposition "...M "effectively knows" what spin of the electron is."[19] (Albert 1992, p. 120) We should note that Albert prefaces this claim with "Let's make up a name for all that..." (Ibid.), so it's not completely clear how to interpret *M*'s 'knowledge' here. However, if 'effective knowledge' *is* somehow claimed to be some kind of knowledge, this claim cannot be correct for two reasons: (1) the 'knowledge' attributed to *M* in this instance is actually a false belief, and as such, can not be *any* form of knowledge, and (2) we have now seen that when in a superposition, *M* ends up without a single belief of any kind about the spin of the electron, and so could not have knowledge about the spin. If an agent such as *M* has no knowledge, Albert's later development of what he calls 'self-measurement' may not work for the bare theory. The reason is that the notion of self measurement for the bare theory requires that an observer like *M effectively knows* the spin of the electron when the outcome is a superposition (Ibid, p. 183). Since as we have seen, *M* can have no knowledge whatsoever about the spin in such circumstances (and even according to Albert will actually have a false belief), *M* will lack *any* knowledge in the process of self-measurement.[20]

[19]Here I use M for h and "spin" for "color".

[20]See Albert and Putnam (1995), and Monton (1998), for further discussion of self-measurement in no-collapse theories. Note that the no-collapse theories they consider (modal theories, for example) all require that "something extra needs to be added" to the quantum state description in order for this self-measurement to occur.(Albert and

6. Disjunctive Experiences

Another peculiar aspect when considering bare theory is the nature of what might be called disjunctive experiences. Here,

> ...a proponent of the bare theory ... would not say that M would determinately believe that she had recorded x-spin up, nor would she say that she would believe that she had recorded x-spin down; rather, she would say that M would determinately believe that she had recorded x-spin up or x-spin down. One might call the experience leading to this disjunctive belief a disjunctive experience.[21] (Barrett 1999, pp. 110-111)

There are three issues to consider here. First, it is not clear what could serve as M's determinate belief in this instance. In analyzing the superpositional state $|\Psi'\rangle$ of the M+S system, the only available belief eigenstates (either introspective or perceptual) are always distinct from one another, and reside in the separate components of the superposition. Further, as we have seen, none of these eigenstates can be factored out of the superposition $|\Psi'\rangle$ to serve as M's determinate belief that she had recorded x-spin up or x-spin down. So there is no single belief state (such as the belief $|D\rangle$ considered above) formed with a disjunctive content like that just proposed.

Second, M's perceptual and introspective beliefs are fine-grained. Recall that the false belief in question is an introspective belief, and so M must be *introspecting* this disjunctive content. Given the fine-grainedness of belief, an introspective belief with the content "x-spin up *or* x-spin down" is distinct from an introspective belief with the content "x-spin up", and both of

Putnam 1992, p. 18; Monton 1998, p. 308) The notion of *adding* something to get an outcome is antithetical to the bare theory, which instead strips quantum mechanics down only to its basic postulates (minus even the collapse postulate), and adds nothing else whatsoever. Since nothing is added to the bare theory (such as the value states of modal theories), the self-measurement issues raised by these other accounts will not occur for an observer in the bare theory.

[21] I have substituted "she" for "he" in this quote.

these are distinct from an introspective belief with the content "x-spin down".[22] The Accuracy Principle insures that the content of any introspective belief in either component of the superposition tracks the content of the perceptual belief eigenstate in that component, which itself tracks the electron spin in that component. Further, these eigenstate relations within each component have evolved deterministically from the Schrödinger equation. Indeed, even the 'disjunctive' introspective contents considered earlier contain contents – "UP" in one component of the superposition and "DOWN" on the other component – [(UP $\vee$ DOWN?: UP) and (UP $\vee$ DOWN?: DOWN)] – which confirm which component of the superposition they reside in, so they are not unresolved disjunctions like the one given in the quote above. So, asking M the content of her introspective eigenstate would yield the fine-grained contents "UP $\vee$ DOWN?: UP" in one component of $|\Psi'\rangle$ and "UP $\vee$ DOWN?: DOWN" in the other component of $|\Psi'\rangle$. Thus, querying M's introspective eigenstates when she is pondering the disjunctive question will not produce a single measurable output. If this belief content is to somehow emerge from a superposition, then linearity requires that this content appear as the introspective content in both components of the superposition. But the fine-grainedness of the contents of the eigenstates in both components of the superposition, as governed by the Accuracy Principle, does not allow there to be a common content for the introspective states.

Which leads immediately into the third issue. As G.E. Moore painstakingly showed over a century ago, introspective and perceptual mental states are *transparent*. That is, when one introspects a perception, one's awareness is of the content of the perceptual state introspected, not the perceptual state itself. Thus any introspective states are drawn to the contents of the state being introspected: introspection reveals no further content than the content of the perceptual state being introspected. (Moore 1903, Dretske 1995, Tye 1995) The belief eigenstates in both components of the superposition contain contents specific to that component ("+" or "–"), but

[22]Each of these introspective beliefs would have a distinct vehicle (eigenstate), content, and causal role. See Dretske (1988) and (1995), and Kim (2010).

never a content of a disjunctive belief with a content of the pure form "x-spin up *or* x-spin down." The common output that is provided is not the content of an introspective, or any other belief state; it remains a common *output* of introspective states which themselves have different contents. No mechanism has been provided by the bare theorist to show that an introspective belief with the required content has been formed. In order to do this, the bare theorist would need to explicate the vehicle, content, and causal role (including its origin) for such a belief state. And that is a tall order that I don't believe the bare theorist has yet provided, because neither component of the superposition contains, nor could produce, such a disjunctive content. Instead, by the Accuracy Principle and by transparency, each introspective and belief content perfectly tracks the pointer result in its component.

7. Concluding the bare theory

According to Albert and Barrett, the bare theory is worth examining because it is the simplest no-collapse theory, and as such, its characteristics will be shared at some level by all other no-collapse theories. Its simplicity also makes the bare theory a good place to start in understanding what features might need to be modified or added to achieve an acceptable no-collapse theory. Albert even calls the bare theory "an amazingly cool idea", and an intriguing way to interpret Everett's theory: "...*this*" he says, "is the idea that it strikes me as interesting to read into Everett's paper." (Albert 1999, p. 124)

However, the bare theory's simplicity also leads to glaring problems, which Albert and Barrett themselves recognize.[23] For starters, how could any pure states required for measuring, and observing, an electron's spin, as represented in equation (6), ever even occur given the evolution of states according the bare theory? Nearby objects would not be limited to maintaining any single trajectory a discreet distance (say) from the experiment, but instead, given the linear

[23] See Albert (1992), Barrett (1999), and also Bub, Clifton and Monton (1998).

dynamics, would evolve into having a superposition of momenta, some of which would result in these objects becoming entangled with the states of the observer, the measuring device, and the electron in the experiment. Even more worrisome, the past histories of the observer, measuring device, and electron, and objects in their vicinity would surely have already created many more such entanglements, complicating matters drastically. The picture the bare theory provides us, therefore, is one of escalating entanglement of nearby objects, which cascades over time to a morass of entangled states which make any hope of conducting an experiment with a determinate outcome futile.

But the difficulties do not stop there. Implementing the bare theory would call into question the very nature of observation and belief, leading one to ask how an observer such as M could ever come to be in a *determinate* perceptual state of observing the detector in the first place, or introspecting that result, or even reporting on its status. Applying the bare theory apparently yields a world without determinate beliefs and reports, leaving no room for sentient beings as we understand them: beings like *us*.

In addition, we have seen that claims of the bare theory regarding the beliefs of an observer of a superposition fall short. To begin, when the observer M answers "Yes" to a question about her perceptual beliefs of the measurement, it does not follow that M would therefore falsely believe that she knows what the result of the measurement is. One reason is that M does not actually have the introspective belief in question; M has no single introspective state that emerges from the superposition with a singular content, including a disjunctive content. For such a single content to emerge from a superposition without such a common belief, linearity would require that the introspective states in both components of the superposition have the same contents; but both the Accuracy Principle and transparency instead insure that each introspective eigenstate has a distinct fine-grained content, which precludes the required single-content result from occurring in both components of a superposition. In addition, M's spoken utterances

originate from output states in Broca's area, which is dedicated to unconscious linguistic processing, not to producing conscious introspections or occurrent perceptions, or indeed any kind of belief. These utterances are outputs of belief states, and not beliefs themselves, and so cannot be false beliefs, since they cannot be intentional in the ways that beliefs are known to be. This phenomenon of producing a common output from a superposition, which lacks the fine-grainedness of the representational states composing the superposition, was shown to also be possible in other systems via the example of a spin detector. This example shows that when a common output can be elicited from a spin detector in a superpositional state, the output lacks the intentional properties required for a misrepresentation about pointer position.

With all these difficulties, what characteristic, aside from its simplicity, makes the bare theory worth considering? As we have seen in this paper, it is the claim by Albert and Barrett that, under the bare theory, observers of superpositions will have false beliefs about their own mental states, and so have "...the illusion of a perfectly ordinary, fully determinate measurement result when there isn't one." (Barrett 2020, p. 148) The promise of observers with false beliefs about their own observations of experiments is indeed fascinating and potentially instructive for no-collapse theories. But there are two problems with this claim as we have now seen. First, it is highly improbable that M, the measuring device, and the electron would start in a pure state like equation (6) without already being entangled with other objects in their vicinity, or even if such a pure state did obtain, that the state could evolve to a superposition like $|\Psi'\rangle$ without being entangled with other nearby objects. But second, even if the experiment did somehow evolve as advertised, a careful analysis of the belief states of the observer under the bare theory shows that M lacks the alleged introspective state altogether, and so she lacks the false belief as claimed. This result leaves the bare theory much less interesting. The promise of observers with false beliefs about their own observations of experiments has been removed, leaving a bare theory that is arguably even more implausible than before.

Acknowledgements

I would like to thank Jeff Barrett for extensive discussions on these topics, and for generously spending time helping me think more clearly through the difficult issues posed by the bare theory. Any mistakes in this paper are purely my own, however. I also thank Reed Guy, John Perry, David Chalmers, Steven Pinker, and Angela Friederici for very helpful discussions and insights on the issues discussed here. Thanks also to Harvey Brown for many fascinating and illuminating lunchtime discussions at Oxford on these ideas. I also thank the anonymous reviewers for their very helpful comments and suggestions.

References

Albert, David (1992), Quantum Mechanics and Experience. Harvard: Harvard University Press.

Barrett, Jeffrey (2020), The Foundations of Quantum Mechanics. Oxford: Oxford University Press.

Barrett, Jeffrey (1999), The Quantum Mechanics of Minds and Worlds. Oxford: Oxford University Press.

Barrett, Jeffrey (1998), "The Bare Theory and How to Fix It." in D. Dieks and P.E. Vermaas (eds.), The Modal Interpretation of Quantum Mechanics, Dordrecht: Kluwer Academic Publishers, 319-336.

Bub, J., Clifton, R. and Monton, B. (1998) "The Bare Theory Has No Clothes" in Healy and Hellman (eds.), Quantum Measurement: Beyond Paradox, Minnesota Studies in the Philosophy of Science, Vol. XVII, Minneapolis: University of Minnesota Press.

Brentano, F. (1874), Psychologie vom Empirischen Standpunkt. Leipzig.

Dehaene, S. (2009), Reading in the Brain. New York, NY:Viking.

Dretske, Fred (1988), Explaining Behavior. Cambridge, MA:MIT Press.

Dretske, Fred (1995), Naturalizing the Mind. Cambridge, MA:MIT Press.

Fleming, S. et al. (2010), "Relating introspective accuracy to individual differences in brain structure." Science, 329(5998):1541-1543.

Frege, G. (1892) "On Sense and Reference" ["Über Sinn und Bedeutung"], Zeitschrift für Philosophie und philosophische Kritik, 100:25–50.

Kim, Jaegwon (2010), Philosophy of Mind, Boulder, CO: Westview Press.

Lee, T.S., et al. (1998), "The role of the primary visual cortex in higher level vision." Vision Research 38:2429-2454.

Monton, Bradley (1998), "Quantum-Mechanical Self-Measurement." in D. Dieks and P.E. Vermaas (eds.), The Modal Interpretation of Quantum Mechanics, Dordrecht: Kluwer Academic Publishers, 308-318.

Moore, G.E. (1903), "The Refutation of Idealism." Mind 12:433-53.

Papineau, David (1987), Reality and Representation. Oxford: Blackwell.

Perry, John (1977), "Frege on Demonstratives." Philosophical Review 86:474-97.

Pinker, S (1994), The Language Instinct. New York: HarperCollins.

Pinker, S (1997), How the Mind Works. New York: W.W. Norton & Co.

Salzman, Britten, and Newsome, "Cortical microstimulation influences perceptual judgements of motion direction." Nature, Vol. 346, No. 6280, pp. 174- 177, 12th July, 1990.

Salzman, et al., "Microstimulation in Visual Area MT: Effects on Direction Discrimination Performance." The Journal of Neuroscience, June 1992, 72(6): 2331-2355.

Seymour, K.J., et al. (2016), "The Representation of Color across the Human Visual Cortex: Distinguishing Chromatic Signals Contributing to Object Form Versus Surface Color." Cerebral Cortex, 26:1997-2005.

Tanaka K, et al. (2003), "Columns for Complex Visual Object Features in the Inferotemporal Cortex: Clustering of Cells with Similar but Slightly Different Stimulus Selectivities."

Cerebral Cortex, 13:90-99.

Tanifuji M., et al. (2001) “Horizontal intrinsic connections as the anatomical basis for the functional columns in the macaque inferior temporal cortex.” Soc Neurosci Abstracts 27:1633.

Tononi, G. (2012), Integrated information theory of consciousness: an updated account. Archives Italiennes de Biologie, 150:290-326.

Tye, Michael (1995), Ten Problems of Consciousness. Cambridge, MA: MIT Press.

Tye, Michael (2000), Consciousness, Color, and Content. Cambridge, MA: MIT Press.

Tye, Michael (2009), Consciousness Revisited. Cambridge, MA: MIT Press.

Zeki, Semir (1993), A Vision of the Brain. Oxford: Blackwell Scientific Publications.