nature communications

Matters arising

https://doi.org/10.1038/s41467-024-47170-2

Quantum mechanical rules for observed observers and the consistency of quantum theory

Received: 15 May 2022

Accepted: 22 March 2024

Published online: 09 April 2024

Check for updates

Alexios P. Polychronakos [1,2]

ARISING FROM D. Frauchiger & R. Renner *Nature Communications* https://doi.org/10.1038/s41467-018-05739-8 (2018)

The interpretation of quantum mechanics in the context of measurements, and concepts such as state "collapse," have troubled physicists since the inception of quantum theory. Pushed to their logical extreme, such issues become entangled (in the colloquial sense) with questions of consciousness, reality, etc. Bell's theorem eliminated possible ways out of this tangle via any (halfway reasonable) classical underlying theory, and the success of quantum mechanics forces us to adopt it as a fundamental theory and face the logical consequences.

The crux of the matter is the privileged role of the observer in quantum theory. Every interpretation of quantum mechanics is formulated in terms of what an observer expects to see or measure (any other formulation would be unscientific). This dichotomy between the observer and (quantum) system raises some obvious questions: are observers subject to the laws of quantum mechanics? (Obvious answer: yes; otherwise the theory would be openly incomplete.) And, who is an observer? (Plausible and possibly incomplete answer: any sufficiently complicated macroscopic system.)

Accepting that quantum theory applies to observers, the issue of what happens to them if they are themselves observed by another (super?)observer becomes important. The question goes back to Schrödinger and his unfortunate cat, and was sharpened by a treatise of Wigner[1] which promoted the cat to another conscious observer, in what is nowadays called the "Wigner's friend" setup. Still, Wigner's thesis neither resolved the question of measurements on observers nor identified any inconsistencies in quantum mechanics and its traditional interpretation.

Recently, Frauchiger and Renner (FR) proposed a thought experiment[2] built on a variant of the "Wigner's friend" setup paralleling a construction by Hardy[3] that, upon application of the conventional rules of quantum mechanics, leads to a contradiction, thus casting doubt on the logical consistency of quantum theory when its application is extended to observers themselves. The FR argument crucially relies on a situation in which an observer is "measured" with respect to a linear superposition of macroscopically distinct states. Then, the application of conventional Born rules and consistency between observers lead to the contradiction. As expected, the arguments and conclusion of FR have become the object of much commentary and debate, and various explanations, remediations, and (often sharply worded) criticism have been offered (Refs. 4–10 is a small, incomplete sample).

In this note, I point out that quantum mechanics requires (in fact, implies) that when such macroscopic measurements happen on observers, then these observers cannot use the standard Born rules of quantum mechanics to predict the results of measurements that will be completed after they suffer such a measurement. In effect, this extends the quantum mechanical mantra "measurements disturb the system" to also apply to observers. Communication of information between such observers is similarly affected.

The argument presented here is applicable to any processes where such measurements are performed and to any deduction based on the predictions of observers for an event after they have suffered such a measurement, implying that conclusions based upon such deductions are not warranted. I will, however, give a more detailed explanation of how the argument applies to the FR thought experiment, which served as the original motivation for this work, and demonstrate that the completion of the quantum mechanical rules proposed here eliminates the inconsistency. I stress that this is not a "refutation" of FR's argument, which solidly relies on a set of assumptions and remains logically valid. It rather complements FR's work by making a (required by quantum mechanics, I argue) modification in their assumptions, which lifts the inconsistency.

Basic approach

I start by declaring that the approach taken in this work relies on strictly unitary evolution of states and on the standard definition of measurement as entanglement between observer and observed system[11]. Unitary evolution is the only one compatible with relativistic quantum mechanics and quantum field theory, and, upon proper interpretation, can account for all observed phenomena[12].

In this approach, there is no fundamental concept of state collapse, and the ensuing certainty of observers about their observation outcomes is encoded in the entanglement between their state and the state of the measured system. After the measurement, the full state

[1]Physics Department, the City College of New York, New York, NY 10031, USA. [2]The Graduate Center, CUNY, New York, NY 10016, USA.
e-mail: apolychronakos@ccny.cuny.edu

becomes a superposition of orthogonal components, each consisting of the state of the observer having observed a particular outcome entangled with the eigenstate of the measured system in that outcome. The full system still represents an observer being certain about the outcome of the measurement, as each orthogonal component shares this property (we may say that the full state is an eigenstate of the "certainty operator").

This approach is essentially equivalent to a "many worlds" interpretation, as each orthogonal component of the state can be considered as a different "branch" of the universe. This interpretation is consistent but not strictly needed: under normal circumstances the orthogonal components are 'superselected'; that is, no transition between them can occur, and thus no physical process can reveal the presence of other branches to an observer in one of them. Other branches are, therefore, epistemologically irrelevant. However, measurements of observers in linear combinations of macroscopic states ("cat" measurements) explicitly induce transitions between the branches, lifting the superselection property and rendering the many worlds interpretation less useful (one may say that branches of the world recombine and mix).

At any rate, the unitary evolution approach is incompatible with other alternative interpretations, such as QBism, and I will have nothing to say about such interpretations.

The basic argument

The argument will be formulated in terms of pure states, but can easily be extended to ensembles of states (density matrices).

Consider the Wigner's friend situation in its simplest: a system consisting of a spin-half S and two observers A and B (whom, since they engage in types of measurements in which Alice and Bob never did, I prefer to think of as Alex and Barbara). A can perform measurements on S, but B can also perform measurements on A, and with respect to states that are superpositions of distinct macroscopic (cognitive) states of A. I will call such states "cat" states (a standard term), and such measurements "cat" measurements.

Initially, the system is in a pure unentangled state $|S\rangle|A\rangle|\mathrm{B}\rangle$ (tensor products $\otimes$ are understood). We work in the Schrödinger representation and assume, for simplicity, that states evolve only when they interact. The process we consider is represented by the state evolution shown below ($|\uparrow\rangle$ and $|\downarrow\rangle$ are the standard z-axis spin eigenstates):

$$\text{Initial state} \qquad \frac{1}{\sqrt{2}}(|\uparrow\rangle+|\downarrow\rangle)|A\rangle|\mathrm{B}\rangle \tag{1}$$

$$\begin{aligned} A \text{ measures spin in } z \text{ axis } &\Rightarrow \frac{1}{\sqrt{2}}(|\uparrow\rangle|U\rangle+|\downarrow\rangle|D\rangle)|\mathrm{B}\rangle \\ &= \frac{1}{\sqrt{8}}\{|\uparrow\rangle[(|U\rangle+|D\rangle)+(|U\rangle-|D\rangle)]+|\downarrow\rangle[(|U\rangle+|D\rangle)-(|U\rangle-|D\rangle)]\}|\mathrm{B}\rangle \end{aligned} \tag{2}$$

$$\begin{aligned} \mathrm{B} \text{ measures } A \text{ in cat state} \Rightarrow \frac{1}{\sqrt{8}}\{&|\uparrow\rangle[(|U\rangle+|D\rangle)|\mathrm{Y}\rangle+(|U\rangle-|D\rangle)|\mathrm{N}\rangle] \\ &+|\downarrow\rangle[(|U\rangle+|D\rangle)|\mathrm{Y}\rangle-(|U\rangle-|D\rangle)|\mathrm{N}\rangle]\} \\ =\frac{1}{\sqrt{8}}&|U\rangle(|\uparrow\rangle|\mathrm{Y}\rangle+|\uparrow\rangle|\mathrm{N}\rangle+|\downarrow\rangle|\mathrm{Y}\rangle-|\downarrow\rangle|\mathrm{N}\rangle) \\ +\frac{1}{\sqrt{8}}&|D\rangle(|\uparrow\rangle|\mathrm{Y}\rangle-|\uparrow\rangle|\mathrm{N}\rangle+|\downarrow\rangle|\mathrm{Y}\rangle+|\downarrow\rangle|\mathrm{N}\rangle) \end{aligned} \tag{3}$$

Initially, the spin is set to the state $|\rightarrow\rangle=(|\uparrow\rangle+|\downarrow\rangle)/\sqrt{2}$. At some time, observer A measures the spin in the z-axis. After the measurement, the states of A and S become entangled, with state $|U\rangle$, representing A having observed an up-spin, entangled with $|\uparrow\rangle$, and state $|D\rangle$, representing A having observed a down-spin, entangled with $|\downarrow\rangle$. After that, the spin is left alone and is not touched by anyone. If A were to measure the spin again, U would definitely find it to be up and D would definitely find it to be down.

At a later time, observer B performs a cat measurement on A. Specifically, B checks if observer A is in the cat state $|U\rangle+|D\rangle$, entangling a state $|\mathrm{Y}\rangle$ representing B having given the answer Yes with the state $|U\rangle+|D\rangle$, and a state $|\mathrm{N}\rangle$ of B having given the answer No with the orthogonal state $|U\rangle-|D\rangle$. The final state is as in (3).

Assume, now, that observer A measures the spin again. The results will be either up or down, irrespective of the value observed previously. If A had originally found the spin to be up, he now has a 50% chance of finding it down. And yet nobody had touched the spin! What has happened is that the observer himself was touched and measured, in a dramatic way that altered the entanglement of his cognitive state with the observed state of the spin.

The lesson we draw from this is:

Observation 1: Observers cannot in general apply the standard Born probability rules if they themselves will be subject to cat measurements.

Observers, of course, do not know the full state of the universe, and often not even the full state of their environment. In general, they know the state of part of their system and update this knowledge as they gather information from measurements they perform or interactions with other observers. This is so, in particular, in the original Wigner's friend setup. For situations not containing cat measurements, deductions based on such partial states are consistent with deductions based on the full state of the system, differing only in the degree of their predictability. Crucially, this is not the case in situations involving cat measurements, and this is the essence of Observation 1 above. To make this explicit, we analyze the situation of eq. (3) in the context of the states perceived by each observer.

Assume, for concreteness, that A and B know nothing initially about the state of the system. A is only aware of the presence of the spin, while B is only aware of the presence of A (their respective measured systems) and, of course, both know their own state. The initial states assumed by each observer are

for A	for B
$\|A\rangle\|S'\rangle$	$\|\mathrm{B}\rangle\|A'\rangle$

(4)

where $|S'\rangle$ and $|A'\rangle$ are generic unknown states for the spin and A. After A performs the measurement of the spin, the updated states are

for A	for B
$\|U\rangle\|\uparrow\rangle$, if up was observed	$\|\mathrm{B}\rangle\|A'\rangle$
$\|D\rangle\|\downarrow\rangle$, if down was observed	

(5)

A can conclude at this point that, if the spin remains undisturbed and he performs his spin measurement again, the probabilities of the outcomes based on his present state are 100% to find the spin up if it was up before, and 100% down if it was down before. After B measures A, the updated states are

for A	for B
$\|U\rangle\|\uparrow\rangle$, if up was observed	$\|\mathrm{Y}\rangle(\|U\rangle+\|D\rangle)/\sqrt{2}$, if Yes was observed
$\|D\rangle\|\downarrow\rangle$, if down was observed	$\|\mathrm{N}\rangle(\|U\rangle-\|D\rangle)/\sqrt{2}$, if No was observed

(6)

Finally, A performs his second measurement of the spin and the updated states are

for A	for B
$\|UU\rangle\|\uparrow\rangle$, if up & then up observed	$\|\mathrm{Y}\rangle(\|U\rangle+\|D\rangle)/\sqrt{2}$, if Yes observed
$\|UD\rangle\|\downarrow\rangle$, if up & then down observed	$\|\mathrm{N}\rangle(\|U\rangle-\|D\rangle)/\sqrt{2}$, if No observed
$\|DU\rangle\|\uparrow\rangle$, if down & then up observed	
$\|DD\rangle\|\downarrow\rangle$, if down & then down observed	

(7)

The two middle outcomes in A's state should not have occurred according to his predictions based on his states at (6). Yet they do

occur, according to the full unitary evolution of the system, and violate A's predictions. A might be tempted to conclude that the spin was disturbed, but this is not a justified conclusion: A could have made sure that the spin was isolated and protected from external influences. The only conclusion that A can draw, then, is that his application of Born rules provided unreliable results.

Are A's unreliable predictions due to his brain having somehow been "scrambled" by the cat measurement? Is A even aware that he has been cat measured? In fact, I would argue that neither is true: in a "clean" cat measurement (involving the minimal measuring operator) the thought process of A is not disturbed. This will be demonstrated later, when the execution and feasibility of cat measurements are examined. At any rate, the effects of a cat measurement on the observer's conscious state and the full details of quantum (cat) vs. classical meddlings with the observer's mind are open to interpretation and might be an issue worth exploring in the future.

Could perhaps observer A modify his application of quantum mechanical rules to account for measurements that he knows will happen to him? Sadly, in general no. To do so, A should know the exact state of the full system before he performs any measurements, as well as the precise measurement that will be performed on him afterwards. With anything short of this full information, A can make no reliable predictions, even probabilistic ones.

As a demonstration, consider that A has no knowledge of the spin state before he measures it, but knows of the presence of B and what exactly she will do to him after he touches the spin. Assuming that A measures the spin and finds it to be up, all that he can deduce is that the state of the spin is now $|\uparrow\rangle$ and the total state is $|\uparrow\rangle|\bar{U}\rangle|\mathrm{B}\rangle$ (with $|\bar{U}\rangle$ the state where A has observed the spin up and knows the measurement to which he will be subjected afterwards, contrasted to state $|U\rangle$ without that knowledge, and similarly for $|\bar{D}\rangle$). Accounting for the upcoming measurement on him, A can deduce the evolution of state

$$A \text{ has measured spin up and deduces state to be} \quad \Rightarrow \quad |\uparrow\rangle|\bar{U}\rangle|\mathrm{B}\rangle \tag{8}$$

$$A \text{ deduces state to become after his cat measurement} \Rightarrow \frac{1}{2}|\uparrow\rangle\left[(|\bar{U}\rangle + |\bar{D}\rangle)|\mathrm{Y}\rangle + (|\bar{U}\rangle - |\bar{D}\rangle)|\mathrm{N}\rangle\right] \tag{9}$$

A can predict that he can observe the spin to be up after he has observed it to be down, but clearly missed the possibility that he can observe the spin to be down after he has observed it to be up. If the initial state was as in (3), and A measured the spin to be up and then concluded that the spin will be measured to be up later on, as implied by (8), he would have 50% probability to be wrong. The lesson we draw is:

Observation 2: Observers cannot in general modify Born rules to fully account for cat measurements on themselves without prior knowledge of the state of the full system and its later evolution.

I should stress that there is nothing unusual about cat states like $|U\rangle \pm |D\rangle$ per se: it is only the possibility of directly measuring them that creates issues. By contrast, their indirect measurement (deduction) poses no problems. For example, consider the scenario where B knows the initial state of the full system and the fact that A will measure the spin in the z-basis, but now she does not measure A; instead, she measures the spin in the x-basis $|\leftarrow\rangle$ and $|\rightarrow\rangle$. The corresponding process would be

$$\text{Initial state} \qquad \frac{1}{\sqrt{2}}(|\uparrow\rangle + |\downarrow\rangle)|A\rangle|\mathrm{B}\rangle \tag{10}$$

$$A \text{ measures spin in } z \text{ axis} \Rightarrow \frac{1}{\sqrt{2}}(|\uparrow\rangle|U\rangle + |\downarrow\rangle|D\rangle)|\mathrm{B}\rangle = \frac{1}{2}[|\rightarrow\rangle(|U\rangle + |D\rangle) + |\leftarrow\rangle(|U\rangle - |D\rangle)]|\mathrm{B}\rangle \tag{11}$$

$$\mathrm{B} \text{ measures spin in } x \text{ axis} \Rightarrow \qquad \frac{1}{2}[|\rightarrow\rangle(|U\rangle + |D\rangle)|\mathrm{R}\rangle + |\leftarrow\rangle(|U\rangle - |D\rangle)|\mathrm{L}\rangle] \tag{12}$$

B now knows that if she has seen the spin to point right ($|\rightarrow\rangle$) then A is in the cat state $|U\rangle + |D\rangle$, and similarly if she has seen it point left, so she has indirectly measured A in a cat state (that is, she has deduced by her knowledge of the state of the system and its evolution that A is in a cat state). However, this causes no problems: although now again A has 50% probability to see the spin up or down, irrespective of what he observed before, he is not surprised, since the spin was disturbed by B's measurement. Crucially, A can use information on what B will measure to make reliable predictions about later measurements based on his updated state after he observes the spin, and without knowledge of the full state before he makes a measurement; a repetition of the steps that led to equations ((4)–(7)) would produce the same final outcomes.

Note that the above statements hold generically. In special situations with specific relations between cat and non-cat measurements, and with observers having partial information on what measurements will be performed, some predictability may be salvaged for them. To demonstrate this, consider the generalized situation of eq. (3) in which A is measured in the new orthogonal cat states $|Y\rangle, |N\rangle$ and the spin is in the state $|\chi\rangle$

$$|Y\rangle = a|U\rangle + b|D\rangle, \quad |N\rangle = b|U\rangle - a|D\rangle \tag{13}$$

$$|\chi\rangle = c|\uparrow\rangle + d|\downarrow\rangle \text{ with } a^2 + b^2 = c^2 + d^2 = 1 \tag{14}$$

(By choosing the phases of $|U\rangle$, $|D\rangle$, $|Y\rangle$ and $|N\rangle$ appropriately we can make a and b real and positive, and similarly for c and d by choosing the phases of $|\uparrow\rangle$ and $|\downarrow\rangle$, since we will not measure the spin in any other basis in this setting.) Following the same sequence of measurements as in (3) we obtain the final state

$$\text{Initial state} \quad |A\rangle(c|\uparrow\rangle + d|\downarrow\rangle)|\mathrm{B}\rangle \tag{15}$$

$$\text{Final state} \quad |U\rangle\left(a^2c|\uparrow\rangle|\mathrm{Y}\rangle + b^2c|\uparrow\rangle|\mathrm{N}\rangle + abd|\downarrow\rangle|\mathrm{Y}\rangle - abd|\downarrow\rangle|\mathrm{N}\rangle\right) + |D\rangle\left(abc|\uparrow\rangle|\mathrm{Y}\rangle - abc|\uparrow\rangle|\mathrm{N}\rangle + b^2d|\downarrow\rangle|\mathrm{Y}\rangle + a^2d|\downarrow\rangle|\mathrm{N}\rangle\right) \tag{16}$$

The probabilities for the unexpected outcomes "A measures spin down given that he first measured it up" p_{ud} and "A measures spin up given that he first measured it down" p_{du} are

$$p_{ud} = \frac{2a^2b^2d^2}{c^2 + 2a^2b^2(d^2 - c^2)}, \quad p_{du} = \frac{2a^2b^2c^2}{d^2 + 2a^2b^2(c^2 - d^2)} \tag{17}$$

These probabilities are maximized for $a = b = 1/\sqrt{2}$, the "maximal" cat state, and become $p_{ud} = d^2$, $p_{du} = c^2$, reproducing the result of eq. (3) for $c = d = 1/\sqrt{2}$. The state of maximal uncertainty for A after having measured the spin up, $p_{ud} = p_{uu} = 1/2$, arises for $c = ab\sqrt{2}$, and similarly after having measured the spin down for $d = ab\sqrt{2}$.

Based on any partial information that A may possess on the initial state of the spin and his upcoming cat measurement, A may have some limited predictive power. E.g., if A is informed that he will be measured in the exact same superposition as the spin ($a = c$, $b = d$

or $a = d, b = c$), then he can deduce "After I measure the spin, the probability to find the opposite value in the subsequent measurement is less than 2/3"; if A is informed that he will be measured in a state correlated with the spin state as in $c = ab\sqrt{2}$, then he can deduce "If I measure the spin and find it up, the next measurement will be completely random; if I find it down, the next measurement is at least as likely to find it down as it is to find it up" ($p_{ud} = 1/2, p_{du} \le 1/2$); etc. However, no general rule emerges for estimating probabilities, and in the absence of any information on the initial state of the spin and the upcoming cat measurement, A is completely ignorant about the outcome of his next spin measurement (both p_{ud} and p_{du} range from 0 to 1).

The above situation also highlights the distinction between cat measurements and ordinary measurements. Observers do interact and "measure" each other continuously, but their interactions produce evolutions of their conscious states and not superpositions of macroscopically distinct states. By contrast, the cat measurement of eq. (13) is part of a continuum that interpolates between no cat measurement ($a = 0$ or $b = 0$) and the maximal cat measurement ($a = b = 1/\sqrt{2}$). As the cat measurement degenerates ($a \to 0$ or $b \to 0$) the probabilities of the surprising outcomes p_{ud} and p_{du} go to zero and the standard Born rules are recovered: there is no "discontinuous" loss of predictability. How such a continuous evolution of states can be obtained with a cat measurement will be described later, when the execution and feasibility of cat measurements are examined.

Communication of information

The previous arguments apply to observer A's prediction of experimental outcomes as experienced by himself. It is also useful, and relevant for the FR thought experiment, to examine how his predictions can be used by other observers; that is, how observers can communicate information.

Consider a third observer C (call him Chris) who does not participate in the measurements, nor is he going to be cat-measured himself, but derives conclusions based on information from A. If observer A directly communicates his prediction to C about the value of the spin before he is cat-measured, then clearly C can treat this information as reliable. Such a communication amounts to entangling the cognitive states of A and C, and therefore of C and the state of the system (spin) measured by A. It is, thus, indistinguishable from C measuring the spin himself. The subsequent cat measurement of A affects neither C nor the spin, and in the absence of cat measurements on himself, C can make reliable predictions.

The situation is similar with indirect (deduced) measurements, that is, for states where cognitive states of A and C become entangled as a result of the dynamical evolution of the system without direct communication between them. If C can reliably deduce such an entanglement from his knowledge of the system, he can treat the information deduced from A's measurement (unreliable for A himself) as reliable. A simple example is the evolution of a state involving two entangled spins and observers A, B, and C:

$$\text{Initial state} \qquad \frac{1}{\sqrt{2}}(|\uparrow\rightarrow\rangle + |\downarrow\leftarrow\rangle)|A\rangle|C\rangle|\mathrm{B}\rangle \tag{18}$$

$$A \text{ measures first spin along } z \Rightarrow \quad \frac{1}{\sqrt{2}}(|\uparrow\rightarrow\rangle|U\rangle + |\downarrow\leftarrow\rangle|D\rangle)|C\rangle|\mathrm{B}\rangle \tag{19}$$

$$C \text{ measures second spin along } x \Rightarrow \frac{1}{\sqrt{2}}(|\uparrow\rightarrow\rangle|U\rangle|R\rangle + |\downarrow\leftarrow\rangle|D\rangle|L\rangle)|\mathrm{B}\rangle \tag{20}$$

$$\text{B measures } A \text{ in cat state} \Rightarrow \frac{1}{\sqrt{8}}\{|\uparrow\rightarrow\rangle[(|U\rangle + |D\rangle)|\mathrm{Y}\rangle + (|U\rangle - |D\rangle)|\mathrm{N}\rangle]|R\rangle + |\downarrow\leftarrow\rangle[(|U\rangle + |D\rangle)|\mathrm{Y}\rangle - (|U\rangle - |D\rangle)|\mathrm{N}\rangle]|L\rangle\} \tag{21}$$

Although A and C never directly interact, the knowledge by C that their states are entangled after C's measurement of the second spin is enough for C to correctly predict the result of a measurement of the first spin, even after A is cat-measured. (Note that the last two measurements commute: performing them in the opposite order changes neither the final state nor the deductions of C and B.).

Things become trickier, however, when C is himself going to be cat-measured. The previous conclusions about predicting or communicating results on a later measurement still hold. However, if the measurement in question is a cat measurement that involves himself, C can neither make reliable predictions, nor transmit reliable information, either directly or indirectly.

Direct transmission of information by C is immediately excluded: this would entangle his state with that of another observer, which would disturb the measured system (himself). What is subtler is the fact that even indirect transmission of information, which would not disturb him, is unreliable. To demonstrate this, consider the process involving observers A (in states $|U\rangle$ and $|D\rangle$) and C (in states $|L\rangle$ and $|R\rangle$) in an entangled state, with B measuring them in cat states $|U\rangle + |D\rangle$ and $|L\rangle + |R\rangle$, and with all observers knowing the full initial state of the system. The state evolution is:

$$\text{Initial state} \qquad \frac{1}{\sqrt{3}}(|U\rangle|L\rangle + |D\rangle|L\rangle + |D\rangle|R\rangle)|\mathrm{B}\rangle \tag{22}$$

$$\text{B cat-measures } A \Rightarrow \frac{1}{2\sqrt{3}}[2(|U\rangle + |D\rangle)|\mathrm{Y}\rangle|L\rangle + (|U\rangle + |D\rangle)|\mathrm{Y}\rangle|R\rangle - (|U\rangle - |D\rangle)|\mathrm{N}\rangle|R\rangle] \tag{23}$$

$$\text{B cat-measures } C \Rightarrow \frac{1}{4\sqrt{3}}[3(|U\rangle + |D\rangle)(|L\rangle + |R\rangle)|\mathrm{YY}\rangle + (|U\rangle + |D\rangle)(|L\rangle - |R\rangle)|\mathrm{YN}\rangle - (|U\rangle - |D\rangle)(|L\rangle + |R\rangle)|\mathrm{NY}\rangle + (|U\rangle - |D\rangle)(|L\rangle - |R\rangle)|\mathrm{NN}\rangle] \tag{24}$$

In the initial state, A in state $|D\rangle$ is entangled with state $|L\rangle + |R\rangle$ of C. Since he knows the initial state of the system, he can deduce this entanglement and he predicts the result Yes for the cat-measurement on C. This prediction is invalid for A himself, in view of his later cat measurement, but can be reliably passed to other observers. C in the state $|R\rangle$ is entangled with $|D\rangle$, so C in that state can inherit the conclusion of A in $|D\rangle$ and indirectly conclude that the result of his own cat measurement will be Yes. This prediction is invalid for C himself, in view of his own cat measurement, but could presumably be reliably passed to other observers.

After B cat-measures A, her state $|\mathrm{N}\rangle$ is entangled with state $|R\rangle$ of C, so if indirect transmission of information from C were reliable, B in state $|\mathrm{N}\rangle$ would conclude that a measurement of C would yield Yes. Yet this prediction is invalidated by the last state in (24), which includes the state $|\mathrm{NN}\rangle$ in which B, originally in the state $|\mathrm{N}\rangle$, obtains the result No for the cat measurement of C.

The lesson we draw from the above chain of arguments is:

Observation 3: Observers cannot, in general, relate reliable information to other observers if both observers are going to be subject to cat measurements.

The previous arguments are also relevant to the operational validity of the assumption of "state collapse." Taking, e.g., the setup described in (8), what observer A is doing is essentially state collapse: based on the information that he obtained from his measurement of the spin, he assumes the state to be an eigenstate of this measurement. This is the best that he can do, lacking any independent knowledge of

the full state before making any observations, and that's what we usually do after measurements, and in general we get away with it: cognitive states corresponding to other possible outcomes do not interfere, and results drawn upon the reduced state by an observer in that state are valid. As demonstrated in eqs. ((4)–(7)), cat measurements change that: by mixing macroscopic states of the observer they make alternatives interfere, and wavefunction collapse yields unreliable results. A restatement of the first lesson of the paper would be:

Observation 4: Observers cannot use state collapse if they will be cat-measured.

The above considerations demonstrate that the deduction rules of quantum mechanics do not hold if the observer suffers cat measurements, and need to be supplemented with the condition of absence of such measurements. This lifts the paradox obtained by FR without modifying the essence of quantum mechanics, as I will demonstrate.

FR's paradox

The FR thought experiment involves two agents F and $\bar{F}$ and two Wigner friends W and $\bar{W}$ in a chain of events and measurements. F and $\bar{F}$ can measure two spins (one of them viewed as a "dice"), while W and $\bar{W}$ can perform cat measurements on F and $\bar{F}$ themselves, labeling the results of these measurements "ok" or "fail." This augmented setup is needed to produce a set of conclusions, derived by consistency between the quantum mechanical predictions of the various observers, that lead to a contradiction. The chain of events and conclusions of the various agents and their mutual interrelation in their temporal succession, as in Table 3 in FR's paper, are summarized below ($|h\rangle$ and $|t\rangle$ are the states of the "dice" spin):

0. The initial state of the two spins is $(|h\rangle|\downarrow\rangle + |t\rangle|\downarrow\rangle + |t\rangle|\uparrow\rangle)/\sqrt{3}$.

1. Agent $\bar{F}$ measures the dice spin, finds it to be $|t\rangle$ and becomes certain that agent W will obtain the outcome "fail" at the final measurement of the experiment (statement $\bar{F}^{n:02}$ in FR)

2. Then agent F performs a measurement of the second spin, finds it to be $|\uparrow\rangle$, becomes certain that $\bar{F}$ observed the dice to be $|t\rangle$, and becomes certain that W will obtain the outcome "fail" at the end because agent $\bar{F}$ is certain of this outcome (statement $F^{n:14}$)

3. Then agent $\bar{W}$ performs a cat measurement on $\bar{F}$'s lab, finds her to be in the $\overline{ok}$ state, becomes certain that F observed the spin $|\uparrow\rangle$ and becomes certain that W will obtain the outcome "fail" because agent F is certain of this outcome (statement $\bar{W}^{n:24}$)

4. Finally, agent W becomes certain he will obtain the outcome "fail" because agent $\bar{W}$ is certain of this outcome (statement $W^{n:28}$), but subsequently measures F and obtains the result "ok", leading to a contradiction

Statement $\bar{F}^{n:02}$ by agent $\bar{F}$ is a prediction based on the application of standard Born inference rules on the specific state $|t\rangle$ that $\bar{F}$ obtains after measuring the dice. Each of the remaining statements 2, 3 and 4 relies on the validity of drawing conclusions based on the previous statement.

FR include the standard quantum mechanical inference rule as one of their basic assumptions (Assumption *Q*). In fact, FR used a weaker, non-probabilistic quantum mechanical rule, applicable to eigenstates of the observed quantity, which was sufficient for the prediction of agent $\bar{F}$ and the derivation of their result. I state their assumption below, slightly paraphrased and in Schrödinger language:

Assumption Q: If an agent A has established at time t_0 that a quantum system S is in a state that will evolve at time t into an eigenstate of an observable X with eigenvalue ξ, then agent A can conclude: "I am certain that $X = \xi$ at time t."

With the additional condition implied by the considerations in the present work, this assumption should be modified as:

Assumption Q′: If an agent A has established at time t_0 that a quantum system S is in a state that will evolve at time t into an eigenstate of an observable X with eigenvalue ξ, and if A knows that no cat measurements will be performed on A during the interval (t_0, t), then agent A can conclude: "I am certain that $X = \xi$ at time t."

The other assumption of FR is consistency between the predictions of different observers (Assumption *C*). I state their assumption below, again paraphrased in Schrödinger language:

Assumption C: If an agent A has established at time t_0 that another agent B, reasoning according to quantum mechanics, is certain that an observable X will have the value ξ at time t, then agent A can conclude: "I am certain that $X = \xi$ at time t."

With the additional condition implied by the considerations in the present work, it should be modified as:

Assumption C′: If an agent A has established at time t_0 that another agent B, reasoning according to quantum mechanics, is certain that an observable X will have the value ξ at time t, and if A knows that no cat measurements will be performed on either A or B during the interval $[t_0, t]$, then agent A can conclude: "I am certain that $X = \xi$ at time t."

The third assumption of FR (Assumption S) is that of logical consistency, precluding the derivation of mutually incompatible results, and is not (and should not be!) modified.

With the assumptions thus modified, FR's argument can stumble at a couple of steps: agent $\bar{F}$ draws her conclusion about the measurement output of agent W at the final step of the experiment based on her present state, leading to statement 1. However, this conclusion is invalidated by the fact that $\bar{F}$ will be herself cat-measured before that final step, as per Assumption *Q*.

Still, this is not necessarily fatal for FR's argument, since $\bar{F}$'s conclusion could possibly be communicated reliably to another agent, leading to statement 2. However, this conclusion is invalidated by the fact that F, who receives this conclusion, will also be cat-measured, triggering the caveat of assumption C. From that point, agent F cannot communicate reliable information to any other agent. Statements 3 and 4 cannot be derived, and no contradiction ensues.

Note that the above argument is valid even if agent $\bar{F}$ is "destroyed" after making the prediction and being measured by $\bar{W}$, as scripted in some scenaria, since $\bar{F}$ does not participate in any of the remaining measurements or deductions. Unitarity forbids the destruction of $\bar{F}$ into a universal destroyed state: each orthogonal state of $\bar{F}$ will be destroyed into distinct orthogonal states, which serve as proxies for the undestroyed states of $\bar{F}$ until the end of the thought experiment, replicating essentially the same state evolution.

This analysis highlights the ingenuity of the thought experiment proposed by FR: a contradiction in quantum theory could easily have been obtained by a simple scenario such as the one of equ. (3), with agent *A* making a prediction for the measurement of the spin after he has measured it once, and seing it invalidated in his subsequent measurement after his own cat measurement. However, FR wanted the contradiction to be obtained by an agent not suffering himself a cat measurement, thus requiring an indirect transfer of information. Yet such information transfers, as I argued around equ. (18), are often reliable. A situation with an unreliable transfer of information was needed, necessitating a second cat-measured agent as well as non-cat-measured agents. In fact, the situation in FR's thought experiment exactly parallels the one in equ. (24), with $\bar{F}$ and F playing the role of *A* and *C*, and B subsuming the roles of $\bar{W}$ and W, while the dice and spin serve to produce the appropriate initial entangled state. Overall, FR's setup is useful in sharpening our intuition and alerting us to the limitations of predictions and communication between observers that suffer cat measurements.

Other arguments for lifting FR's paradox have been offered, and the difficulties caused by cat measurements have been highlighted, with statements such as Scott Aronson's witty aphorism "It's hard to think when someone Hadamards your brain"[4], or Lenny Susskind's comment in Renato Renner's seminar[10] about "closed loops" in the many-world interpretation. My arguments sharpen the issue into

precise statements and propose specific modifications of Assumptions *Q* and *C*. They also eliminate the possibility of a general modification of quantum rules (based only on observationally available data) to take into account cat measurements, or at least show that the quantitative rules for such modifications are nontrivial and as yet to be formulated. As I stated early in the paper, I prefer to eschew the many-worlds view as it offers no conceptual advantages in the presence of cat measurements, since the question of "who can branch the world?" is essentially equivalent to "who can collapse wavefunctions?".

Finally, it should be obvious why cat measurements are necessary to produce FR's paradox while classical measurements would not do it. An agent could make a prediction and reliably relate it to another agent before getting confused by a classical "bang on the head," producing no inconsistencies. By contrast, the information deduced from two cat-measured observers can become unreliable as their states are scrambled after their cat measurements, which is a pure quantum effect. This is the essence of eq. (24) and of FR's thought experiment, and this is what the modified Assumption C warns about.

Are cat measurements possible?

The possibility (or suspicion) of cat measurements performed upon ourselves or our experimental apparati would be catastrophic for our ability to usefully apply quantum theory. The success of quantum mechanics in every context where it was applied so far is evidence that such measurements are either physically impossible or of vanishingly small probability. Observers, of course, do interact and "measure" each other continuously, but their interactions are essentially classical, that is, they never create superpositions of macroscopically distinct states.

The von Neuman realization of a cat measurement on A would require coupling the measured system with the momentum of the position operator of the "needle" of the observation apparatus. Such an interaction for the process (3) would be

$$h_I = \lambda\, p\, \Pi \tag{25}$$

with λ a real coupling constant, p the momentum operator dual to the position x of the needle of a measuring apparatus in observer B's lab, and Π an operator with $|U\rangle + |D\rangle$ and $|U\rangle - |D\rangle$ as non-degenerate eigenstates. Up to irrelevant additive and multiplicative constants, such an operator can be expressed as

$$\Pi = |U\rangle\langle D| + |D\rangle\langle U| \tag{26}$$

and would act as

$$\Pi\,|U\rangle = |D\rangle, \quad \Pi\,|D\rangle = |U\rangle \tag{27}$$

That is, Π is an exchange operator that acts on A and changes his state from one where he has observed the spin to be up to one where he has observed it to be down, and vice versa. Applying the Hamiltonian h_I for time t produces the unitary evolution

$$\mathrm{U} = e^{-ih_I t} = \frac{1}{2}e^{-i\lambda t p}(1+\Pi) + \frac{1}{2}e^{i\lambda t p}(1-\Pi) \tag{28}$$

Starting with the initial state $|U\rangle|\mathrm{B}\rangle$, the state at time t woul be

$$\mathrm{U}|U\rangle|\mathrm{B}\rangle = \frac{1}{2}(|U\rangle + |D\rangle)|\mathrm{B}(\lambda t)\rangle + \frac{1}{2}(|U\rangle - |D\rangle)|\mathrm{B}(-\lambda t)\rangle \tag{29}$$

where $|\mathrm{B}(x)\rangle$ represents the state of B with her measuring device needle's position shifted by x (the initial state of B would be $|\mathrm{B}\rangle = |\mathrm{B}(0)\rangle$). If the initial uncertainty in the position of the needle is δ, then after time $T > \delta/\lambda$, B could decide with certainty if the needle moved in the positive or negative direction, and at that time $|\mathrm{B}(\lambda T)\rangle = |\mathrm{Y}\rangle$ and $|\mathrm{B}(-\lambda T)\rangle = |\mathrm{N}\rangle$, leading to the final state

$$\frac{1}{2}(|U\rangle + |D\rangle)|\mathrm{Y}\rangle + \frac{1}{2}(|U\rangle - |D\rangle)|\mathrm{N}\rangle \tag{30}$$

The important fact is that the state in (29) never contains a state of confusion for A; it is always a superposition of $|U\rangle$ and $|D\rangle$. Both states $|U\rangle$ and $|D\rangle$ are undisturbed states of clear certainty about the value of the spin (up or down) and the full state at all times is an eigenstates of A's "certainty operator." This justifies the statement that A would feel absolutely nothing during a "clean" measurement such as the one above and would not even be aware that he is cat-measured. It also demonstrates the asymmetry in the situation: only A suffers the action of exchange operators. Ironically, A never leaves a state of certainty, while B goes through a continuous set of states of uncertainty $|\mathrm{B}(\lambda t)\rangle$ until she reaches her final state of certainty about the outcome of the measurement.

Are cat measurements such as the one above physically realizable? In fact, the physics of performing cat measurements is prohibitive. Exchange operators are strongly nonlocal (essentially effecting "teleportation") and hard to realize, even in the simplest of systems. For example, the parity operator P reflecting the position and momentum of a particle on the line, can be realized as

$$P = \exp\left[i\frac{\pi}{2}\left(ax^2 + \frac{p^2}{a\hbar^2} - 1\right)\right],\ Px = -xP,\ Pp = -pP \tag{31}$$

with a a nonzero real constant (despite appearances, P is both Hermitian and unitary). This is a highly unphysical operator, involving an infinite sequence of local operators (upon Taylor-expanding the exponential). Physical interactions are local, and no finite sequence of them would reproduce P.

The realization of Π would similarly involve nonlocal operators acting on the macroscopically large number of particles making up observer A, and in a highly coordinated pattern, pushing it outside the realm of physical possibilities. Interactions between observers, however intense or even violent, are a collection of local individual interactions and will never reproduce Π. Even a reasonable approximation of Π would need to involve an exceedingly long sequence of operations whose execution would require a time likely exceeding the lifetime of the universe.

Nevertheless, the question of whether cat measurements are in principle realizable is an interesting one and remains essentially open. I offered some arguments why such measurements would be practically impossible, but at the conceptual level it would be desirable to have a proof of their full impossibility, perhaps involving locality, relativity, quantum field theory (which does not even contain strictly factorizable, unentangled states of finite energy) or other physical principles. In fact, making the question a meaningful one would require thermodynamics to enter the argument at some level. Just as there is no sharp distinction between "small" (quantum) and "large" (classical) systems, what constitutes an observer, and thus what is a cat measurement, is equally fuzzy. It is often argued (or conjectured) that the arrow of time and the manifestation of consciousness are related to entropy flow. In that case, a physically meaningful definition of cat measurements would necessarily involve large systems out of equilibrium. The physical realization of operators like Π could then possibly be excluded by entropic considerations.

In conclusion, quantum mechanics is alive and well, still challenging us to understand it to our intellectual and emotional satisfaction. If cat measurements can be ruled out, quantum mechanics will become more reliably predictive. If not, cat measurements will remain in our intellectual playground and may lead to interesting and weird effects, and possibly new insights, although not to inconsistencies. I am biased for the former, but otherwise remain agnostic.

Data availability

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

References

1. Wigner, E.P. "Remarks on the mind-body question," Symmetries and Reflections, Ch. 13, 171–184 (Indiana University Press, 1967).
2. Frauchiger, D. & Renner, R. Quantum theory cannot consistently describe the use of itself. *Nat. Commun.* **9**, 3711 (2018).
3. Hardy, L. Quantum Mechanics, Local realistic Theories, and Lorentz-Invariant Realistic Theories. *Phys. Rev. Lett.* **69** (1992).
4. Aaronson, S. blogpost entry of Sep 25, (2018).
5. Araujo, M. blogpost entry of Oct 24, (2018).
6. Lazarovici, D. & Hubert, M. How Quantum Mechanics can consistently describe the use of itself. *Sci. Comm.* **9**, 470 (2019).
7. Kastner, R. E. Unitary-Only Quantum Theory Cannot Consistently Describe the Use of Itself: On the Frauchiger-Renner Paradox. *Found. Phys.* **50**, 441 (2020).
8. Bong, K.-W. et al. A strong no-go theorem on the Wigner's friend paradox. *Nat. Phys.* **16**, 1199 (2020).
9. Guérin, P. A., Baumann, V., Del Sonato, F. & Brukner, Č. A no-go theorem for the persistent reality of Wigner's friend's perception. *Commun. Phys.* **4**, 93 (2021).
10. Renner, R. 3/1/2021 SITP seminar and audience comments.
11. Hepp, K. Quantum theory of measurement and macroscopic observables. *Helv. Phys. Act.* **45**, 237 (1972).
12. Hardy, L. Nonlocality for two particles without inequalities for almost all entangled states.*Phys. Rev. Lett.* **71**, 1665 (1993).

Acknowledgements

I would like to thank Stuart Samuel for interesting me in the work of Frauchiger & Renner and for relating his own, different, explanation of the paradox, and Parameswaran Nair for a critical assessment of my argument. I am especially thankful to Renato Renner for a useful correspondence and for sharing his insights. This work was supported by NSF under grant NSF-PHY-2112729 and by PSC-CUNY grants 65109-00 53 and 6D136-00 02.

Author contributions

A.P. is the sole contributor to this work and is fully responsible for its contents.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Alexios P. Polychronakos.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.