What is it like to be unitarily reversed?

Peter W. Evans

University of Queensland

Abstract

There has been in recent years a huge surge of interest in the so-called extended Wigner's friend scenario (EWFS). In short, a series of theorems (with some variation in detail) puts pressure on the ability of different agents in the scenario to account for each of the others' measured outcomes: the outcomes cannot be assigned single well-defined values while also satisfying other reasonable physical assumptions. These theorems have been interpreted as showing that there can be no absolute, third-person, 'God's eye' description of our reality. The focus of this paper is the strongest of these no-go theorems, the 'local friendliness' theorem of Bong et al. (2020), which gives earnest consideration to the possibility of a measurement that unitarily reverses an entire lab system, including a conscious agent, thereby erasing the agent's memory. The purpose of this paper is to begin the philosophical conversation regarding key questions concerning this process: Are the events in the lab merely 'erased', or do they in some sense not exist at all? What would it be like to be unitarily reversed? Should an agent care about any experiences they have inside the lab before they are reversed? This analysis employs a parallel case of memory erasure, to which this case can be contrasted, arising in the context of drug-induced amnesia as a result of administering anaesthesia during medical procedures (Carbonell, 2014). I argue that the consequences of unitarily reversing an agent are much more dramatic than simply memory erasure—the set of events themselves, and the personal timeline of the agent, leave no record at all inside or outside the lab. I consider the ramifications of this for the picture of reality that arises from the EWFS.

Keywords: Quantum foundations; Extended Wigner's friend scenario; Local friendliness; Wigner bubble

1 Introduction

There has been in recent years a huge surge of interest in the so-called extended Wigner's friend scenario (EWFS). The EWFS consists of two agents—the 'friends'—who are each isolated in a lab where they measure one half of a bipartite entangled quantum system, and two further agents outside each lab—the 'superobservers'—who perform some quantum operation on their friends inside the lab. First introduced by Frauchiger and Renner (2018) to form the core of a no-go theorem, the scenario has become integral to a series of no-go theorems: in particular, those of Brukner (2018) and Bong et al. (2020). In short, the theorems (with some variation in detail) put pressure on the ability of different agents in the scenario to account for each of the others' measured outcomes: given the right combination of each of the four agents' measurements, and that each agent can be certain that their own outcome is definite, then the four outcomes cannot be assigned single well-defined values while also satisfying other reasonable physical assumptions. These theorems have been interpreted as showing that there can be no absolute, third-person, 'God's eye' description of our reality.

The focus of this present paper is an intriguing feature of the strongest of these no-go theorems, the 'local friendliness' theorem of (Bong et al., 2020). To establish their theorem, Bong et al. consider the following experimental setup for the EWFS. The two superobservers can choose one from a series of measurements to perform on the system inside the lab, including the friend. The first such choice is simply to open the lab and ask the friend what their outcome is. But the authors specifically consider one other type of measurement that the superobservers can perform: a unitary reversal of the entire lab system, which erases the outcome of the friend's measurement of the entangled quantum system, followed by a measurement of the local half of the entangled quantum system in a different basis to the one used by the friend. As the authors point out, however, this unitary reversal of the lab system also results in the friend having their memory erased.[1] However, very little attention is paid to the implications of this claim (although one of the authors explores an interpretation of such unitary reversal in some depth (Cavalcanti, 2021)). For instance, is it physically possible to unitarily reverse an experiencing agent? Is it right to think of the friend's measurement outcome as being erased, or rather as having not existed at all? What exactly would it feel like to be unitarily reversed? Should this agent care about any experiences they have inside the lab before they are reversed?

The purpose of this paper, then, is to begin the conversation concerning some of these implications. In particular, I consider the friend's experience of their personal timeline. This analysis employs a parallel case of memory erasure, to which this case can be contrasted, arising in the context of drug-induced amnesia as a result of administering anaesthesia during medical procedures (Carbonell, 2014). After outlining this case in §2, I flesh out in §3 the detail of the scenario that would lead to the unitary reversal of a conscious agent. I then consider what it would be like to be unitarily reversed in §4, and whether an agent should care about the events that took place before reversal in §5, drawing further on the parallel case of drug-induced amnesia. I argue that the consequences of unitarily reversing an agent are much more dramatic than simply memory erasure—the set of events themselves, and the personal timeline of the agent, leave no record at all inside or outside the lab. I consider in §6 the ramifications of this for the picture of reality that arises from the EWFS.

2 Drug-induced amnesia and personal identity

In her paper "Amnesia, Anesthesia, and Warranted Fear", Vanessa Carbonell considers the ramifications for personal identity of psychological discontinuities that arise in situations where patients experience drug-induced amnesia as a result of conscious anaesthetic sedation during medical proce-

[1] This particular measurement setup is chosen by Bong et al. as it is useful for simplifying the statistics and the subsequent explanation of violations of the local friendliness inequalities. Of course, it is possible that other kinds of measurements could feature in such a proof, such as a projection on an entangled basis of the lab system (Brukner, 2018). However, as Allard Guérin et al. (2021) have shown, even considering more general measurements on the lab system, "treating the memory of [the friend's] measurement outcome as having a value throughout the experiment is in conflict with important features of quantum mechanics". This result of course generalises to any record of the measurement outcome inside the lab. Thus it would seem that regardless of the particular measurement setup, instances of erasure of measurement records may be more ubiquitous than simply the measurement setup that Bong et al. consider.

dures.[2] Carbonell's focus is the issue of fear and anxiety, and how these may be, on the one hand, assuaged by loss of memory from anaesthesia—"don't worry, you wont remember a thing"—and, on the other hand, exacerbated by the very same amnesia, particularly for patients who are undergoing invasive or potentially painful medical procedures, who lack the contextual memory of the procedure, and who therefore fill in the gaps in their memory with imagined or exaggerated fears. She argues that a patient's fears *should not* be assuaged by the prospect of not remembering some invasive or unpleasant procedure, as the patient will after all be conscious enough to experience this unpleasantness (that is why drug-induced amnesia might be thought desirable in the first place: precisely to forget the unpleasantness). Thus, in addressing whether the knowledge that amnesia will be induced as a result of conscious anaesthetic sedation during a medical procedure should reduce the anxiety of the pre-operative patient over the prospect of an invasive, painful, or unpleasant procedure, Carbonell concludes that the pre-operative patient should rightly fear any pain that the peri-operative patient experiences, but should not fear the prospect of post-operative trauma from any such experience.

Importantly for our purposes here, to argue this point Carbonell makes use of Parfit's (1984) psychological continuity criterion of personal identity and considers how patients should consider their sense of self over time in situations where drug-induced amnesia has disrupted the continuity of their personal identity. According to Parfit's criterion, personal identity requires psychological continuity, which is brought about by direct causal connections between a person's current psychological state and their past psychological states—that is, their memories (Parfit, 1984, p.206). This process of identification of different psychological states as belonging to a single individual thus extends across time, with each moment of experience representing a time slice of the individual that is causally influenced by past time slices, and causally influences future time slices.

In the case of our patient undergoing an experience of drug-induced amnesia as a result of conscious anaesthetic sedation during a medical procedure, this psychological continuity is disrupted, as the memories and experiences from the period of time spanned by the amnesia are no longer accessible to the post-operation patient. However, as Carbonell (2014, p.248) puts it, after the patient re-emerges from her drug-induced amnesia (emphasis in original):

> she is not (due to the amnesia) directly and strongly psychologically connected with the person who was in pain. But she *is* directly connected to herself *before* the operation. And *that* person has a strong connection with the person during the operation, since she was conscious the entire time. Thus the post-operative patient is connected to the pre-operative patient directly, and connected to the suffering patient *indirectly*, via the pre-operative patient.

[2] There is, of course, a long and storied history to the connection between personal identity and psychological continuity. While a deep engagement with this tradition is well beyond the scope of this project, I provide here some of the major punctuation points in that story. The discussion originates in (Locke, 1694) with the idea that memory is constitutive of personal identity, followed by the objection that memory ultimately presupposes identity, first by (Butler, 1736) and then by (Reid, 1785). The connection between memory and personal identity is defended against this circularity in a contemporary setting by both (Shoemaker, 1970) and (Parfit, 1984), while (Schechtman, 1990) instead defends the circularity objection. With respect to amnesia and psychological discontinuity and its relation to personal identity, (Brennan, 1985) supports their compatibility, while (Schechtman, 2005) argues that personal identity requires a more practical 'self-understanding' component, in addition to simply memories, which underpins an ongoing 'narrative' about an individual's life, even in cases of psychological discontinuity; see also (Klein and Nichols, 2012) for a discussion of personal identity in a neurological case study of memory loss. The current project relies on Carbonell's analysis of drug-induced amnesia due to the significant parallels between that case and the unitary reversal of an agent in the EWFS. (Although, another case from the personal identity literature with clear parallels is Elga's (2000) 'Sleeping Beauty' problem. This would be interesting future work to pursue.)

Thus the key point here is that psychological continuity is preserved between the pre-operative patient, A_i, the peri-operative patient, A_o, and the post-operative patient, A_f, since there remains a transitive strong connectedness relation between A_f and A_i, and then between A_i and A_o. All this is to say, quite obviously, that there are no grounds on Parfit's account of personal identity for the post-operative patient to claim that the peri-operative patient is in some sense not her. Yes, the post-operative patient's psychological states are not directly causally influenced by the psychological states of the peri-operative patient, but they are directly causally influenced by the psychological states of the pre-operative patient, as are the psychological states of the peri-operative patient.

Carbonell's reasoning to argue her main result—that the pre-operative patient should rightly fear any pain that the peri-operative patient experiences—is that the relationship between the pre-operative patient and the peri-operative patient is the same regardless of whether amnesia is induced or not (Carbonell, 2014, fn.23):

> how strongly interested the pre-operative patient is in the peri-operative patient's wellbeing. . . depends on how much psychological unity holds between them. But there is plenty of psychological unity between the pre-operative patient and her peri-operative, pained self. So she should be quite concerned about that pain. . . And as far as this question [how should we anticipate the procedure?] is concerned, the answer seems to be: just as you would if you wouldn't be getting amnestic drugs. . .

The similarities between this case of drug-induced amnesia and its connection to personal identity and the case of an agent being unitarily reversed should be starting to become clearer. In the next section I explain the local friendliness theorem in more depth, before spelling out these similarities more explicitly in §4. Notably, though, there are some important differences, too, and I hope to use these differences to draw out the potential ramifications of the EWFS for our picture of reality.

3 Unitary reversal of a conscious agent

The thought that a living being could possibly be made to undergo coherent quantum evolution arises within the first few years after the theory of quantum mechanics was developed. Famously, Schrödinger (1935) imagines a cat in a steel chamber whose fate is entwined with the decay of a small amount of radioactive material. While the import of the thought experiment was to point out the ludicrous consequences of what Schrödinger then dubbed 'entanglement', the possibility of coherently evolving a living being has since caught the public imagination. The possibility of the coherent quantum evolution of a *conscious* agent became a more serious consideration after Wigner (1961) introduced his own thought experiment in which he considers his friend interacting with a quantum system, and himself interacting with the entire 'friend-plus-quantum-system' arrangement. We now know this as the 'Wigner's friend' thought experiment, and it forms the basis of the extended Wigner's friend scenario (EWFS).

As I mentioned in the opening passages, the EWFS consists of the following experimental setup: two agents—the 'friends'—are each isolated in a lab where they measure one half of a bipartite entangled quantum system, and two further agents outside each lab—the superobservers—perform some quantum operation on their friends inside the lab. According to the no-go theorem of Bong et al. (2020), when a set of reasonable assumptions hold, there are formal constraints on the possible correlations between the agents' outcomes, which they call the 'local-friendliness' (LF) inequalities. The three assumptions are: (i) observable events are conditionally independent of spacelike separated measurement choices; (ii) the agents in the experiment can freely choose their actions (there are no 'cosmic conspiracies' constraining their behaviour); and (iii) events that are observed by any agent are real single events that are not relative to anything or anyone (Bong et al., 2020, p.1201). Thus, if an appropriate quantum system violates the LF inequalities, then one of these three assumptions must be inadmissible (a result they point out is strictly stronger than Bell's (1966) result).

Importantly for our purposes here, one of the key mechanisms in the setup is that the superobservers can choose one from a series of measurements to perform on the system inside the lab, one of which is a unitary reversal of the entire lab system, friend and all. Thus, more than just imagining the coherent quantum evolution of a living being, as Schrödinger and Wigner did, Bong et al. consider the unitary evolution of a conscious agent from some initial state, such as when the friend closes the lab door, to some intermediary state, presumably after the friend has made their measurement, and then back again to the initial state of the lab system.[3] The authors point out that this will serve to erase both the outcome of the friend's measurement of the entangled quantum system and also the friend's memory of anything that occurred inside the lab.

Despite the fact that a setup such as this incorporating actual conscious agents remains for the time being hypothetical, Bong et al. are interested in providing proof-of-principle experimental evidence for quantum violations of the LF inequalities, where in their experiment the role of the two 'friends' are played by two distinct photon paths. And they do indeed find quantum violations of the LF inequalities. However, since the quantum information inherent in the two photon paths representing each friend illuminates the implications of the local friendliness theorem only in so far as a photon can be considered an 'observer', and cannot illuminate the original Wigner's friend thought experiment, we are constrained to the realm of the hypothetical at least for now. However, this does not mean that this sort of experiment will not be possible in the future, and so I contend that the philosophical consequences of this setup deserve proper attention.[4]

Before exploring some of these philosophical consequences, let us consider a couple of issues with the EWFS that are worthy of note. Firstly, one might argue that the coherent unitary evolution of a macroscopic system is simply impossible. On the one hand, we might find that there is some mechanism of objective collapse at the relevant microscopic scale, which would rule out macroscopic coherence of the sort required in the EWFS. Similarly, it might be that, since it is exceedingly difficult to properly isolate a lab from any gravitational field, such effects ensure decoherence is ubiquitous, and so no macroscopic coherent state is possible. Both Bong et al. and Brukner explicitly note that such objective collapse theories would circumvent the impetus of their respective theorems.[5] In fact, Bong et al. imply (perhaps as a kind of *reductio*) that the undesirability of violations of local friendliness might require some sort of radical revision of quantum mechanics along the lines of objective collapse proposals. Let us assume for the sake of this philosophical exploration that

[3] Strictly, Bong et al. consider the unitary evolution of an *observer*; the above claim holds in so far as consciousness is required for a system to be an observer.

[4] For comparison, it was roughly 50 years between Bell's theorem and the loophole-free experimental tests (Hensen et al., 2015).

[5] Although, see Wiseman et al. (2023) for further discussion on this point in the context of a human-level AI simulation on a quantum computer.

macroscopic coherence is possible.

A second issue of interest concerns the process of 'reversal' of the conscious agent. For any coherent quantum system, the state of the system can be represented as a vector in Hilbert space, and the evolution of the system as the rotation of that vector through the space. If we think of the coherent evolution of the lab system in this way, such that the interaction of the friend plus quantum system is simply the rotation of a vector in Hilbert space, then it is not necessarily the case that we need to reverse the system to put it back in its initial state. If the initial conditions are known, we could restore the system back to its initial state by simply evolving the system to the desired initial state. However, if the initial conditions are unknown, but the evolution operators are known, we can restore the system back to its initial state by reversing each of the evolution operators, regardless of whether we know that state or not. These mechanisms are formally equivalent.

With these preliminary issues to one side, let us consider the parallels between unitary reversal of a conscious agent and drug-induced amnesia as a result of conscious anaesthetic sedation during medical procedures.

4 What is it like to be unitarily reversed?

Consider the perspective of the friend as she enters the lab and closes the door—call the state of the friend and lab system at this point F_{b_1}. If we adopt the psychological continuity criterion of personal identity as we did above, it is clear that the friend inside the lab is strongly psychologically continuous with the friend who arrived to the lab earlier in the day, F_i: the psychological state of the friend at F_{b_1} is appropriately causally connected to the psychological state of the friend at F_i; these friend states are uncontroversially the same person. The friend then proceeds to measure her half of the shared entangled bipartite quantum system, and record the result (perhaps by writing it down in her notepad)—call the state of the friend and lab system at this point F_{b_2}. The superobserver outside the lab now enacts the reversal process (either by evolving the system back to its initial state, or sequentially reversing each evolution operator). This restores the friend back to F_{b_1}. The superobserver then opens the lab door, and the friend exits and begins interacting with the outside world (for instance, by chatting with the superobserver about her experiences)—call the state of the friend and lab system at this point F_f.

If we consider again the psychological criterion of personal identity, we can step though a parallel argument to the one we developed above in connection with the drug-induced amnesia patient. As above, psychological continuity is preserved between the friend before entering the lab, F_i, the friend inside the lab, F_{b_1} and F_{b_2}, and the friend after exiting the lab, F_f, since there remains a transitive strong connectedness relation between F_f and both F_i and F_{b_1}, and then between both F_i and F_{b_1} and the later lab state F_{b_2}. There is a clear sense in which the psychological state of the friend at F_{b_2} is directly causally influenced by that at F_{b_1}, which is directly causally influenced by that at F_i, and which both directly causally influence the psychological state of the friend at F_f, even though, due to having her memory erased during the unitary reversal, the psychological continuity between F_{b_1} and F_f has been disrupted as there is no direct causal influence between them.

However, there is also a sense in which the case of the unitarily reversed friend is quite different from the drug-induced amnesia case. Consider the psychological states of the friend between F_{b_1} and F_{b_2}. These states no longer leave a causal influence on F_f, just as was the case for the amnesia patient. But something more dramatic has happened: it is not just the psychological states of the friend between F_{b_1} and F_{b_2} that has been erased, but *records of* all the events inside the lab have been erased, too. In the drug-induced amnesia case, even though psychological continuity was disrupted, there was a fact of the matter about the events that took place that the patient could not remember; these events left a record in the surrounding environment (the memory of the operating doctor, the hospital records, etc.). But, by construction, none of the events inside the lab have left any record in the surrounding environment, as the evolution was unitary. Depending on how one might interpret the quantum formalism, it is unclear whether there is a fact of the matter about the events that took place inside the lab. The entire lab system is in state F_{b_1} just after the friend closes the lab door, and is in precisely that state again the moment before the superobserver opens the door to let the friend out.

Cavalcanti (2021) describes the set of events that took place inside the lab as occurring in a 'Wigner bubble'. It is worth emphasising here exactly what it means for the restored state F_{b_1} to be identical to the state just after the friend closes the lab door. It is not just the psychological state of the friend that is the same at the two temporal ends of the Wigner bubble, it is the friend's complete physical state—the friend will not have aged a second. The principal reason for this is that there is no entropy gradient underpinning the ageing process inside the lab that can differentiate the moment the friend enters the lab and the moment she exits. But what happens between those times inside the lab is an open matter. On the one hand, one might argue that the evolution of the lab system is entirely unitary, and so must be isolated from decoherence with the environment outside the lab. In so far as unitary evolution is not entropy increasing, then there are no records generated and there is no entropy gradient throughout the evolution inside the lab. On the other hand, one might argue that there is indeed a local entropy gradient internal to the lab system, and this gradient underpins the creation of internal records such as the outcome of the friend's local measurement, especially if one relies on a dynamical von Neumann model of measurement (Mello, 2014). On this latter view, the friend first evolves with a local entropy gradient in some sense 'aligned' with the gradient outside the lab, but then when the lab is unitarily reversed, the local entropy gradient also reverses, and one might argue that the friend 'experiences' the same events in the reverse temporal direction. We will revisit these points in just a moment.

However, there is a tension on this latter view between the unitary evolution of the lab system and the idea that local records are created inside the Wigner bubble, underpinned by some entropy gradient. Even if we were to take a von Neumann model of measurement, wherein the process of measurement consists of the dynamical evolution of the target system interacting with some measurement probe, ultimate projection is still a fundamental element of the model of measurement. One could argue that such measurement events are definitively irreversible. If this is the case, then it just would not be possible to unitarily reverse a lab system within which our friend makes some measurement on a quantum system. Or, alternately, the friend simply could not make a proper measurement if the evolution of the lab were unitary. This tension undermines the possibility of the orthodox Wigner's friend scenario, let alone the EWFS.

5 Should agents care what happens in a Wigner bubble?

So far we have considered what the claim of unitary reversal of a conscious agent might entail, and we have employed Carbonell's framework to provide a neat entry point for thinking about the associated philosophical implications. However, we have not considered as yet what this might mean for the reality of any events that occur inside a Wigner bubble. One suggestion for whether we should consider events in a Wigner bubble as real comes from Cavalcanti (2021), who provides a sustained philosophical analysis of the kind of constraints that the EWFS places on interpretations of quantum mechanics. Employing a QBist (Caves et al., 2002) analysis of the EWFS, Cavalcanti sets out a framework for understanding the meaningfulness of the probabilities assigned to the different events by different agents in the EWFS. Cavalcanti argues that despite the fact that the events that take place in a Wigner bubble are not in fact events at all from the superobserver's perspective, "this does not amount to... a rejection of the (relative) existence of the friend's perspective" (Cavalcanti, 2021, p.28). Even though "[t]he question of which outcome was observed is not (pragmatically) meaningful to [the superobserver]", it is certainly the case that "it is meaningful from the friend's perspective" (Cavalcanti, 2021, p.29). There is thus an inherent tension between the two agents' perspectives—the friend and the superobserver—and what they each say about the reality of the events in the Wigner bubble. (This is, of course, precisely the tension that a rejection of assumption (iii) to avoid violating the LF inequalities entails, and so why the EWFS is so interesting.)

While Cavalcanti leaves a detailed discussion of this part of his argument to one side, considering the question of whether the friend *cares* about her future self and the events inside the lab, in comparison to the parallel case of drug-induced amnesia from §2, can provide an interesting angle on Cavalcanti's analysis of this tension. Recall from above that Carbonell concludes that the pre-operative patient should rightly fear any pain that the peri-operative patient experiences, since the relationship between the pre-operative patient and the peri-operative patient is the same regardless of whether amnesia is induced or not. But just as the pre-operative patient should care about what happens in the operation, so should the friend care what happens to her in the Wigner bubble. This is despite the fact that the events inside the bubble effectively do not exist according to any observer outside the bubble.

Let us raise the stakes a little. Let us imagine an EWFS where the friend is to enter an isolated lab and, dependent upon the result of her measurement of the quantum system, she is to be subject to torture.[6] Once the torture has been administered for some set time interval, the whole lab system will be unitarily reversed such that for any agent outside the lab none of the events inside the lab have any reality. The friend is told of this plan when she arrives before the experiment begins; it is clear that she should very much care about what happens to her inside the Wigner bubble. The reasoning for this runs precisely parallel to the case of drug-induced amnesia. What matters for whether an agent should care about their potential future self is the psychological unity between them now and their potential future time slice. The psychological relation between the friend before the experiment and the friend inside the lab seems entirely independent of whether the lab system is unitarily reversed or not. Thus it seems the friend would be right to fear the experiment.

If one were to take a QBist perspective on the EWFS, as Cavalcanti does, then it follows that

[6] I thank Eric Cavalcanti for this suggestion.

the probabilities that an agent ascribes to observable events need to be in principle associated with some kind of bet about those events that can be settled. In so far as any such probability assignment is pragmatically meaningful—perhaps two agents inside a Wigner bubble are betting against each other about what they will observe—then one might infer that this pragmatic meaningfulness is underpinned by the reality of the events inside the bubble. I suggest that the above considerations are not inconsistent with Cavalcanti's QBist analysis of the Wigner bubble: the friend's probability assignments inside the Wigner bubble regarding, say, the length and severity of her torture are pragmatically meaningful *to the friend* only in so far as those events are underpinned by the reality inside the bubble. By the same token, the concern that the friend feels for her future self is likewise sufficient on this QBist reading to give those events a kind of reality *to the friend* that they do not have for agents outside the bubble. Since the EWFS is increasingly interpreted as showing that there can be no absolute, third-person, 'God's eye' description of our reality, this inherent tension appears to be simply an expression of this consequence.

But this is only one interpretation-specific analysis of reality inside a Wigner bubble, and it rests on Cavalcanti's inferred connection between a QBist reading of the pragmatic meaningfulness of probability assignments and the reality of the events that underpin those assignments. My view, however, is that there are deeper tensions to surmount in understanding the reality of events in a Wigner bubble. Let us consider in the next section one of these implications in particular.

6 What is an agent without leaving a trace?

As mentioned above, the agents involved in the EWFS at the core of the local friendliness theorem are at the moment hypothetical. So one must take the preceding analysis in the spirit of conceptual exploration rather than a declaration of definite physical results. Of course, one logical possibility in response to this analysis is that, in concert with an implied suggestion from Bong et al. (2020, p.1199), the above considerations amount to an argument in favour of rejecting the local friendliness assumptions and embracing an interpretation of quantum mechanics that rejects the universal applicability of the theory. But if we are to take the most obvious consequence of the local friendliness theorem—that the reality of observed events are relative to the observer—seriously, then we will need to tackle some of the above conceptual difficulties sooner or later. And the hope is that some of the above analysis can provide a starting point for this discussion.

The key take-away from this paper is that working through the consequences of unitarily reversing a conscious agent illuminates some of the worries that arise from the claim that there can be no absolute, third-person, 'God's eye' description of our reality. However, concerns regarding what it is like to be 'reversed' are not unique to Wigner's friend experiments, nor even to quantum mechanics. A similar issue arises in the context of the Poincaré recurrence theorem in classical physics (Poincaré, 1890), which states that a closed, conservative dynamical system with a bounded state space will return to a state arbitrarily close to its initial condition within some finite timescale. Thus, one could consider a hypothetical agent as above, strictly isolated from their environment (again, as above), and imagine that within some astronomically large but finite time frame the system will return to its initial state.[7]

[7] Some level of caution is required here, since not all parts of the phase volume will return to the initial configuration

There is a clear overlap between this classical case and the quantum case discussed above. If a classical agent were able to exist long enough within the isolated system for the system to return to its initial state, many of the considerations above would apply in this case also. Most significantly, none of the events in the isolated classical system, by construction, would have left any records either within the system itself, including the memory of the agent, nor outside in the surrounding environment. What is unorthodox about this scenario is that an agent is not typically a conservative dynamical system, but rather a highly dissipative system, and so is not usually considered a candidate system for the Poincarè recurrence theorem—but this is of course the same situation as faced by Wigner's friend, and generates the dissonance between the unitary evolution of the friend inside the lab and the idea that the creation of local records underpinned by some entropy gradient is possible.

There is one significant difference in the way we might think about the reality of both the classical and quantum isolated systems. We typically think that the intervening states of the classical system are perfectly real, determinate, spatiotemporally-localised physical states despite the fact they leave no record either within or outside the system. And perhaps there is a temptation to interpret the Wigner's friend scenario in the same manner. However, and this is the importance of the local friendliness theorem and the EWFS in the context of this discussion, so long as we assume an absence of both nonlocal influences and cosmic conspiracies (as we noted in §3) then we cannot avail ourselves of the same kind of classical reality underpinning the quantum description of the lab system and still expect that that reality remains compatible with the observations of quantum mechanics. How then should we understand the 'events' inside the Wigner bubble on this view? I do not wish to prescribe a definitive answer to this question here, but I do flag that all answers seem to come with substantial metaphysical baggage.[8]

There is one final consideration, however, that might be worth further attention. Recall the point above that an agent who is unitarily reversed would exit the Wigner bubble having not aged from the moment the bubble was entered. I suggested that this is due to the fact that, since the evolution forwards and backwards through the states of the system inside the bubble is unitary, there is no entropy gradient underpinning any ageing process for the agent. Anti-ageing aside, there is another reason why an entropy gradient is important to an agent. It seems reasonable to define an agent as a being that can act on and affect the world around them (Schlosser, 2019). By establishing a boundary around the friend (the bubble) within which the friend is evolved unitarily (and potentially reversed), and in doing so eradicating an entropy gradient with which the friend could leave records in the environment outside the bubble, we thus remove the ability of the friend to affect any part of

at the same time (Myrvold, 2021). As such, since the agent is a complex dynamical system, the whole agent system may not reappear in its initial state simultaneously. That would make a strange kind of agent indeed.

[8] There are, of course, many further deep ramifications for metaphysics as a consequence of this result. For instance, it has become somewhat of an orthodoxy in the philosophy of time, particularly as a result of the philosophical implications of both the special and general theory of relativity, that all past and future times are equally as real as the present—this is the so-called B-theory of time. It seems a key implicature of the B-theory that there is indeed a 'God's eye' description of all the events in the world. The most plausible consequences of the EWFS looks to seriously disrupt this orthodoxy, for which there is some precedent in the literature on the philosophy of time (Fine, 2005; Iaquinto and Torrengo, 2022). In this context, there is perhaps an interesting connection to be drawn between an interpretation of the events inside the Wigner bubble 'having not existed at all', and decidedly 'anti-B-theory' conceptions of time travel and 'changing the past' (see, for instance, (van Inwagen, 2010; Effingham, 2021) and for a counterpoint to these arguments (Baron, 2017)). I am personally inclined towards arguments along the lines of (Ismael, 2023) as a path to progress on these consequences of the EWFS. I thank an anonymous reviewer for suggesting this connection.

the world outside the bubble, and so we potentially remove any sense of agency from the friend as well. Does the friend still get to count as an agent?

One might argue that the friend herself will draw the boundary between herself and the world where any agent would normally draw that boundary, and so she would be choosing her actions based on the impression that she were freely able to affect the world. But there are two countervailing factors here. The first is that any such influence on the world would be confined to the Wigner bubble within which the friend found herself. As such, if those interactions with the world are then reversed, then there is a sense in which the process of reversing the bubble takes away the 'agency' behind those actions. The second countervailing factor is related to the issues raised at the end §4 above, concerning the tension between unitary evolution and measurement. If the stipulation that the lab system evolves unitarily rules out the possibility of anything that we might ordinarily call a 'measurement', then this would definitively rule out the possibility of the friend performing any action at all inside the Wigner bubble that could qualify them as an 'agent' acting in the world—even in their restricted bubble world.

This may seem a line of thought with no real practical implications. However, Bong et al. (2020) and Cavalcanti (2021) set up the possibility for future tests of quantum violations of LF inequalities by suggesting that the role of the friend could be played by a strong AI in a universal quantum computer (Bong et al., 2020, p.1203):

> If universal quantum computation and strong AI are both physically possible, it should be possible to realize quantum coherent simulations of an observer and its (virtual) environment, and realize an extended Wigner's friend experiment. . . Towards the goal of challenging the LF no-go theorem, experiments can test agents of increasing complexity; an experimental violation of LF inequalities with a given class of physical systems as 'friends' implies that either the LF assumptions are false or that class of friends is not an 'observer'.

This raises a series of questions that are perhaps empirical, and push the boundaries of what we might consider an agent. If a strong AI on a universal quantum computer acts in a virtual environment, is its claim to agency pinned to whether it leaves records in its virtual environment, and so whether it exploits a virtual entropy gradient? Or is its claim to agency pinned to whether it leaves records in the *real* environment of the quantum computer (which it most certainly would do as the quantum computer processed the information enabling the simulation to take place)?[9] But now consider the following variation of this thought: if the strong AI enters a virtual Wigner bubble, and so is evolved unitarily and then reversed, not only does it plausibly leave no record in the virtual environment, but since the quantum computer on which the simulation is taking place unitarily evolved the simulation, the quantum computer also leaves no record in its *real* environment.[10] By the same token as we asked the question above, does the strong AI still get to count as an agent in this setup? By the looks of things, an answer to this question in the more straightforward (but more hypothetical) case above should point to an answer in this virtual case. And in so far as this virtual case is the most promising

[9] These questions are of course not peculiar to quantum computers simulating a quantum environment, but arise for classical simulations and virtual worlds as well. In this context, Chalmers (2022) argues that "virtual worlds are as real as an ordinary physical world". Answering the corresponding question in the classical case would be highly interesting and likely nontrivial.

[10] The actual situation may be more complex than this. The Hamiltonian being calculated by a quantum computer is time dependent and requires a precisely controlled clock. However, inevitable fluctuations in the period of any such clock will introduce errors into this calculation. As such, a quantum computer will need to be irreversibly connected to a low entropy system that corrects any errors. Whether this amounts to a sufficient 'record' in the real environment is an interesting question, but is beyond the scope of this argument.

proposal for future tests of quantum violations of LF inequalities, the above critique of the notion of agency takes on increased significance.

So while the above considerations are indeed constrained to the hypothetical, one suspects that it will not be long before iterative steps towards a (real or virtual) implementation of the EWFS will begin to emerge. And it is then that the preceding discussion will take on increasing importance. It seems prudent to begin the conversation now.

Acknowledgements

For useful discussion and comments I would like to thank Eric Cavalcanti, Gerard Milburn, Howard Wiseman, and three anonymous referees for EJPS. This work was supported by the Foundational Questions Institute and Fetzer Franklin Fund project “Quantum and consciousness: paths to experiment, and implications for interpretations” (FQXi-RFP-CPW-2019).

References

Allard Guérin P, Baumann V, Del Santo F, Brukner Č (2021) A no-go theorem for the persistent reality of Wigner's friend's perception. Communications Physics 4:93, doi:10.1038/s42005-021-00589-1

Baron S (2017) Back to the Unchanging Past. Pacific Philosophical Quarterly 98(1):129–147, doi:https://doi.org/10.1111/papq.12127

Bell JS (1966) On the Problem of Hidden Variables in Quantum Mechanics. Reviews of Modern Physics 38:447–452, doi:10.1103/RevModPhys.38.447

Bong KW, Utreras-Alarcón A, Ghafari F, Liang YC, Tischler N, Cavalcanti EG, Pryde GJ, Wiseman HM (2020) A strong no-go theorem on the Wigner's friend paradox. Nature Physics 16:1199–1205, doi:10.1038/s41567-020-0990-x

Brennan A (1985) Amnesia and Psychological Continuity. Canadian Journal of Philosophy 11:195–209, doi:10.1080/00455091.1985.10715896

Brukner Č (2018) A No-Go Theorem for Observer-Independent Facts. Entropy 20(5):350, doi:10.3390/e20050350

Butler J (1736) Of Personal Identity. In: The Analogy of Religion, John James and Paul Knapton, chap First Appendix, reprinted in Perry (1975), pp. 99–105

Carbonell V (2014) Amnesia, anesthesia, and warranted fear. Bioethics 28(5):245–254, doi:10.1111/j.1467-8519.2012.01995.x

Cavalcanti EG (2021) The View from a Wigner Bubble. Foundations of Physics 51(2):39, doi:10.1007/s10701-021-00417-0

Caves CM, Fuchs CA, Schack R (2002) Quantum probabilities as Bayesian probabilities. Physical Review A 65(2):022305, doi:10.1103/PhysRevA.65.022305

Chalmers DJ (2022) Reality+: Virtual Worlds and the Problems of Philosophy. Penguin Books

Effingham N (2021) Vacillating time: a metaphysics for time travel and Geachianism. Synthese 199:7159–7180, doi:10.1007/s11229-021-03108-5

Elga A (2000) Self-Locating Belief and the Sleeping Beauty Problem. Analysis 60(2):143–147, doi:10.1093/analys/60.2.143

Fine K (2005) Modality and Tense: Philosophical Papers. Clarendon Press, Oxford

Frauchiger D, Renner R (2018) Quantum theory cannot consistently describe the use of itself. Nature Communications 9:3711, doi:10.1038/s41467-018-05739-8

Hensen B, Bernien H, Dréau AE, Reiserer A, Kalb N, Blok MS, Ruitenberg J, Vermeulen RFL, Schouten RN, Abellán C, Amaya W, Pruneri V, Mitchell MW, Markham M, Twitchen DJ, Elkouss D, Wehner S, Taminiau TH, Hanson R (2015) Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. Nature 526:682–686, doi:10.1038/nature15759

Iaquinto S, Torrengo G (2022) Fragmenting Reality: An Essay on Passage, Causality and Time Travel. Bloomsbury Publishing, London

van Inwagen P (2010) Changing the Past. In: Zimmerman DW (ed) Oxford Studies in Metaphysics, Volume 5, Oxford University Press, New York, chap 1, pp 1–28

Ismael J (2023) Rethinking Time and Determinism: What happens to determinism when you take relativity seriously. In: Lestienne R, Harris PA (eds) Time and Science, Volume 1: The Metaphysics of Time and Its Evolution, World Scientific Publishing, chap 5, pp 147–172, doi:10.1142/9781800613737_0005

Klein SB, Nichols S (2012) Memory and the Sense of Personal Identity. Mind 121(483):677–702, doi:10.1093/mind/fzs080

Locke J (1694) Of Identity and Diversity. In: Essay Concerning Human Understanding, Book II, A. and J. Churchil, and S. Manship, chap XXVII, reprinted in Perry (1975), pp. 33–52

Mello PA (2014) The von Neumann model of measurement in quantum mechanics. AIP Conference Proceedings 1575(1):136–165, doi:10.1063/1.4861702

Myrvold WC (2021) Beyond Chance and Credence: A theory of hybrid probabilities. Oxford University Press, USA

Parfit D (1984) Reasons and Persons. Oxford University Press, Oxford

Perry J (1975) Personal Identity. University of California Press, Berkeley

Poincaré H (1890) Sur le problème des trois corps et les équations de la dynamique. Acta Mathematica 13(1):A3–A270, doi:10.1007/BF02392505, translated by Bruce D. Popp in Poincaré (2017)

Poincaré H (2017) The three-body problem and the equations of dynamics: Poincaré's foundational work on dynamical systems theory. Springer, Berkeley

Reid T (1785) Of Identity. In: Essays on the Intellectual Powers of Man, Essay III, J. Bell, chap 4, reprinted in Perry (1975), pp. 107–112

Schechtman M (1990) Personhood and Personal Identity. The Journal of Philosophy 87(2):71–92, doi:10.2307/2026882

Schechtman M (2005) Personal Identity and the Past. Philosophy, Psychiatry, & Psychology 12(1):9–22, doi:10.1353/ppp.2005.0032

Schlosser M (2019) Agency. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, Stanford University, plato.stanford.edu/archives/win2019/entries/agency/

Schrödinger E (1935) Die gegenwärtige Situation in der Quantenmechanik. Naturwissenschaften 23:807–812, doi:10.1007/BF01491891

Shoemaker S (1970) Persons and Their Pasts. American Philosophical Quarterly 7(4):269–285, www.jstor.org/stable/20009360

Wigner EP (1961) Remarks on the Mind-Body Question. In: Good IJ (ed) The Scientist Speculates: An Anthology of Partly-Baked Ideas, Heineman, chap 13, pp 171–184

Wiseman HM, Cavalcanti EG, Rieffel EG (2023) A "thoughtful" Local Friendliness no-go theorem: a prospective experiment with new assumptions to suit. Quantum 7:1112, doi:10.22331/q-2023-09-14-1112