Consciousness and the Collapse of the Wave Function

Kelvin McQueen
Philosophy department, VU University Amsterdam
Department for Physics and Astronomy, Tel Aviv University
k.j.mcqueen@vu.nl

[with David Chalmers]
Philosophy department, New York University
Philosophy department, Australian National University

Purpose of this talk

- Explore a largely unexplored interpretation of quantum mechanics.
  - The consciousness causes collapse interpretation (CI).

- First motivate CI as:
  - A solution to the problem of outcomes.
  - An attempt to stay close to the “face-value” interpretation.

- Then defend and develop CI using:
  - Integrated information theory of consciousness, and
  - Spontaneous collapse dynamics.

Quantum Mechanics: predictions

- A familiar story about QM predictions:
  - Physical systems described by a wave function ψ.
  - Typically, ψ specifies *superpositions* of values for a given property.
  - Typically, ψ evolves via deterministic Schrödinger equation.
  - Measurements yield definite values, determined probabilistically by pre-measurement ψ and Born Rule.
  - Post-measurement ψ updated to reflect definite values.
- For predictions, this story has been tremendously successful!
  - Canonical statement: Von Neumann [1932/1955]

Quantum Mechanics: reality

- A natural thought:
  - The story is predictively successful because the story *describes reality*.
    - Quantum realism
- The "face-value" interpretation:
  - Physical reality fundamentally involves a wave function with a *bipartite* dynamics.
    - Schrödinger equation
      - Linear, deterministic, constantly ongoing.
    - Collapse
      - Nonlinear, nondeterministic, happens only during *measurement*.

Measurement problem

- QM predictions story: widely accepted.
  - Predictive success.
- QM reality story: widely rejected.
  - Measurement problem:
    - Notion of “measurement” is vague, anthropocentric, and inappropriate in a fundamental specification of reality.

- The face-value interpretation is itself intended to solve a deeper problem.
  - The problem of outcomes…

Problem of outcomes

- Propositions (A-C) are incompatible:
  - (A) The state of a physical system is completely specified by the wave function.

    $|\text{Alive}\rangle_{cat}\ \alpha(|\text{Decay}\rangle + |\sim\text{D}\rangle)_{atom}$
  - (B) The wave function evolves only in accord with a linear equation (e.g. Schrödinger equation).

    $\alpha|\text{Dead}\rangle_{cat}\ |\text{Decay}\rangle_{atom} + \alpha|\text{Alive}\rangle_{cat}\ |\sim\text{D}\rangle_{atom}$
  - (C) Measurements have definite singular outcomes.

    $\alpha|\text{Alive}\rangle_{cat}\ |\sim\text{D}\rangle_{atom}$

    - "Interpretations" (quantum theories) are solutions to this problem and are classified in terms of which proposition they reject…

Interpretations as solutions

- (A) The state of a physical system is completely specified by the wave function.
  - Denied by: additional variables theories (e.g. Bohm)
- (B) The wave function evolves only in accord with a linear equation (e.g. the Schrödinger equation).
  - Denied by:
    - Face-value theory (von Neumann)
    - Near-face-value theories
      - **Consciousness-induced collapse** (London & Bauer, Wigner)
    - Spontaneous collapse theories (e.g. GRW, Pearle)
- (C) Measurements have definite singular outcomes.
  - Denied by: many-worlds theory (e.g. Everett)

From Face-value theory to CI

- Motivations for CI:
  - 'Conscious observation' is an analysis of 'measurement'.
  - Consciousness superpositions seem impossible.
- Wigner's (1967) theory:
  - Physical reality fundamentally involves a wave function with a *bipartite* dynamics.
    - Schrödinger equation
      - When there are *no* conscious experiences.
    - Collapse
      - When there *are* conscious experiences.
  - Problem: No formalism offered to model these interactions.
    - Squires (1993): "I am not aware of any model equations with such properties and […] I cannot conceive of how they could be constructed."

Building a formalism

- We need:

- (1) a formal definition of consciousness, which treats consciousness as a (measurable) quantity.

- (2) mathematically precise collapse equations describing consciousness-matter interactions.

But hasn't CI already been refuted?

- Many bad objections have been proposed.
  - It's solipsism
  - It's idealism
  - It's dualism

- But also some important objections e.g.
  - No improvement on face-value theory since 'consciousness' no better than 'measurement'.
  - Relies on mind-matter interaction which we know nothing about.

David Albert's objection to (1)

- 'Conscious' has no precise meaning...
  - "But the trouble here is pretty obvious too: What this "theory" predicts (that is: what "theory" it is) will hinge on the precise meaning of the word conscious; and that word simply doesn't have any absolutely precise meaning in ordinary language." (1992: 83).

- Initial response:
  - Ordinary language is not relevant.
  - Language of consciousness science is.

Peter Kosso's objection to (2)

- "The important interaction between mind and matter is unexplained. There is no clue as to the mechanism by which consciousness affects physical objects and causes the collapse of the state function. The consciousness interpretation does not offer progress since it explains one mysterious phenomenon (the collapse of the state function during measurement) in terms of an equally mysterious phenomenon (the interaction between mind and matter)." (1998, 171).

- Initial response:
  - Perhaps we can make progress by proposing empirically testable mechanisms.

In response

- Response to Albert:
  - Phenomenal consciousness.
  - Integrated information theory of phenomenal consciousness (IIT).

- Response to Kosso/Squires:
  - Combine IIT and spontaneous collapse dynamics.

‘Consciousness’ in neuroscience

- Block (1995) contrasts *access* consciousness with *phenomenal* consciousness.
- A state is ***phenomenally conscious*** iff:
  - ...there is something it is like for the system to be in that state. (Think: what is it like to see red, to feel happy, to have an itch?)
  - Also: what you lose when you fall into dreamless sleep (Tononi).
- Tononi’s *integrated information theory* offers a fundamental theory of consciousness.

Integrated information theory (IIT)

- Conscious experiences are *informationally differentiated* and *integrated*.
  - A given experience rules out possibilities, yielding information.
  - A given experience is unified and integrated.
- So are the neural correlates of consciousness.
  - Cerebral cortex vs cerebellum.
- Integrated information is mathematically quantifiable and measureable.
  - Amount of Integrated information = ϕ ("phi")
- IIT hypothesis: ϕ quantifies phenomenal consciousness.

Response to Albert

- We have: (1) a formal definition of consciousness, which treats consciousness as a measurable quantity.
  - Amount of consciousness measured by ϕ
- Albert's objection therefore not convincing.

- Now recall Kosso's objection...

- We also need:
  - (2) mathematically precise dynamical (collapse) equations that describe the interactions between ordinary physical systems and conscious systems.
    - Let's start with the face-value theory...

The face-value theory

- How close can we stay to the face-value theory?
  - Measurement devices instantiate a "measurement" property - which *cannot superpose*.
  - "Measurement" properties respond to impending superpositions by collapsing the wave function.
    - ***M-properties*** **defined by this functional role.**

- Let ϕ be our m-property and be done with it?
  - No: Quantum Zeno effect.

Zeno effect Vs. Face-value theory

- Quantum Zeno effect:
  - If system S is in eigenstate E of some observable O, and measurements of O are made N times a second, then, the probability that S will be in E after one second tends to one as N tends to infinity.
- Problem for M-properties:
  - M-property evolves by Schrödinger equation when not being superposed *by other systems*.
  - But Schrödinger equation will *immediately* superpose the M-property.
  - So M-property continuously collapses to its initial value.
    - M-property *freezes*.
    - But consciousness evolves.
- Face-value theory refuted?
  - No: explore *approximate m-properties* as in GRW.

GRW theory: collapse dynamics

- At random times particle wave-function Ψ *collapses*:

$$\psi_t(x_1, x_2, \ldots x_N) \rightarrow L_n(x)\psi_t(x_1, x_2, \ldots x_N)$$

- Ψ_t = wave function prior to the collapse; and:

$$L_n(x) = Ae^{-(q_n - x)^2/2r_C^2}$$

- A = constant.
- q_n = position operator for particle *n*.
- r = localization width = 10^{-7}m.
- x = place of collapse or "collapse centre".
  - Probability of x being collapse centre = $|L_n(x)|\Psi_t>|^2$.

GRW theory: overall dynamics

- GRW combine:
  - standard Schrödinger equation
    - Which is ongoing…
- With: $\lambda[L_n(x)]$
- $L_n(x)$ = collapse function from previous slide.
- λ = temporal distribution of collapses = $10^{-16}s^{-1}$.
  - So, each second, an isolated particle has a 10^{-16} probability of *spontaneously* collapsing.
  - For entangled N-particle macro-system: $\lambda_{macro} = N\lambda$.

Spontaneity and *ad hoc*-ness

- GRW's basic idea:
  - An isolated particle collapses, on average, **every hundred million years**.
  - Consequently, a macro-system collapses about every 10^{-7} seconds.

- Every hundred million years, *spontaneously*?
  - Odd property for particles to have! Seems a*d hoc*.

- Face-value and GRW theory may converge on a deeper theory….
  - Removing both QZE and ad hoc-ness?

The Kremnizer-Ranchin (KR) Model

- Recall GRW's modification:

$$\lambda \left[Ae^{-(q_n - x)^2 / 2r_C^2} \right]$$

- KR replace λ so that collapse frequency is a function *f* of the ϕ of the system's density matrix ρ :

$$f\left[\varphi\left(\rho(t)\right)\right]\left[Ae^{-(q_n - x)^2 / g\left[\varphi\left(\rho(t)\right)\right]}\right]$$

- Note that the *form* of the collapse (g) is also an (unspecified) function of ϕ.
  - QZE problem solved?
  - Ad hoc-ness removed?

A modification

- Can we give consciousness a more direct role? Let's replace function f with function j:

$$j\left[\varphi\left(\rho(t)\right)\right]\left[Ae^{-(q_n-x)^2/g\left[\varphi\left(\rho(t)\right)\right]}\right]$$

- Let j be a function of the ϕ *variance* over superposition components.
  - Then consciousness is approximately an m-property.

Experimental tests of CI

- CI is experimentally verifiable:
  - Firstly: calculate integrated information of various interesting physical systems.
    - Mesoscopic computers?
  - Secondly: compare collapse rate of various systems with very different ϕ–values.
    - Requires precise control of the environments of these systems (to avoid decoherence disruptions).
- CI is potentially *easier* to test than GRW since high-ϕ systems need not be large.

The mind-body problem

- Have we answered Kosso's objection?

- We have an empirically testable model for certain causal effects of phenomenal consciousness.
  - Approximate m-property model.
  - Further (philosophical, mathematical & experimental) development may therefore contribute to a solution to the mind-body problem.
    - CI therefore deserves to be taken seriously.

Conclusion

- CI is no worse off than spontaneous collapse theories, and so deserves to be taken as seriously.
  - CI is arguably less ad hoc.
  - CI may be easier to test experimentally.

- CI is an open research program with a clear way to proceed based on:
  - (i) the integrated information theory of consciousness and
  - (ii) spontaneous collapse dynamics.

References

- Albert, D.Z. 1992. *Quantum Mechanics and Experience*. Harvard University Press.
- Barrett, A.B. 2014. An Integration of Integrated Information Theory with Fundamental Physics. *Frontiers in Psychology* 5(63): 1-6.
- Block, N. 1995. On a Confusion about a Function of Consciousness. *Behavioral and Brain Sciences* 18(2): 227-287.
- Griffith, V. 2014. A Principled Infotheoretic ϕ-like Measure. arXiv preprint arXiv:1401.0978.
- Kremnizer, K. & Ranchin, A. 2015. Integrated Information-Induced Quantum Collapse (V2). arxiv.org/abs/1405.0879.
- Kosso, P. 1998. *Appearance and Reality: An Introduction to the Philosophy of Physics*. Oxford University Press.
- Maudlin, T. 1995. Three Measurement Problems. *Topoi* 14: 7-15.
- McQueen, K.J. 2015. Four Tails Problems for Dynamical Collapse Theories. *Studies in the History and Philosophy of Modern Physics* 49: 10-18.
- Tononi, G. 2012. Integrated Information theory of Consciousness: an updated account.
- von Neumann, J. 1955. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press. German original: *Die mathematischen Grundlagen der Quantenmechanik*. Berlin: Springer, 1932.
- Wigner, E.P. 1967. Remarks on the Mind-Body Question. In *Symmetries and Reflections*. Indiana University Press. pp. 171–184.

Appendix 1: Integrated Information Theory (IIT)

Integrated Information theory (IIT)

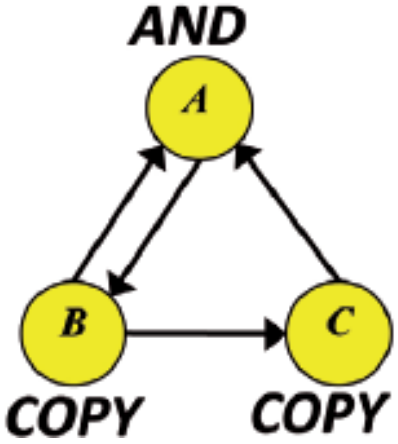
- IIT starting observation: consciousness is highly *differentiated* and highly *integrated*.
  - *Differentiated*: a given conscious state is information-rich by virtue of ruling out alternative possibilities.
    - Experiencing darkness yields knowledge that it's not-dark in all conceivable ways (not-red, not blue etc.).
  - *Integrated*: this information is highly unified.
    - I experience a whole visual scene, not RHS plus LHS.
    - Seeing a red triangle not reducible to seeing colorless shape plus disembodied redness.
- IIT core hypotheses:
  - *Amount of consciousness* in a system corresponds to *amount of integrated information* ϕ in a system.
  - ϕ is a mathematically well-defined measurable quantity.

Neuroscientific support for IIT

- Cerebellum: not relevant for consciousness despite many neurons; cerebral cortex: relevant for consciousness despite few neurons.
  - IIT: neural activity in cerebellum less orchestrated (less ϕ).
- Split brain patients: cutting cortical hemisphere links, splits consciousness in two.
  - IIT: ϕ of whole brain slips below ϕ of each hemisphere.
- Probe the (conscious) cerebral cortex with a pulse of current, cortex responds with widespread (integrated) and differentiated (information rich) reverberating activations and deactivations.
  - But as consciousness fades, cortex response becomes local (less integrated) or global but stereotypical (less information).

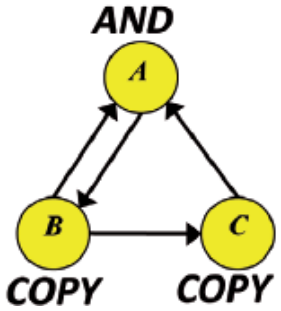
Information

- The following comes from Tononi (2014).
- Consider element A's *mechanism*:
  - Logical AND gate of elements B & C.
  - Turns on if B&C are ON.
  - Turns B on if A is OFF.

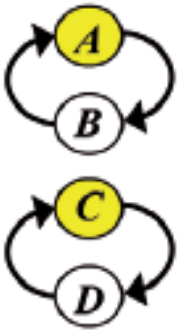


- Let A be ON. Then A's mechanism is:
  - Consistent with only 2 of the 8 possible *past* states of ABC.
  - So, *effective info:* $EI[(ABC)_{past} | A_{present}] = 6$.

Integration



- To calculate ϕ given EI, you must calculate the product of EI of *partitions* e.g.
  - $EI[(AB_{past} | A_{pres}) \times (C_{past} | A_{pres})]$
- Then find the *minimum information partition* MIP that yields the least difference from the whole.
- Then ϕ equals the *difference* between the EI of the whole and the EI of the MIP.
  - $\phi = D[(PAST|PRES) , (\Pi(PAST/PRES)/MIP)]$



Integrated information ϕ

- Clearly, ϕ is difficult to calculate (even for wire diagrams).
  - For a brain, consider the difficulty of finding MIP!
    - But work is being done on simplifying this (e.g. Griffith 2014).
- Key point for our purposes:
  - There is a theoretically motivated measure of consciousness that is in principle possible to calculate for real systems.

Quantum Integrated Information (KR)

- The QII of a density matrix is:

$$\phi(\rho) = \inf(S(\rho||\ {}_{i=1}^{N}\bigoplus Tr_i(\rho)) :\ H \approx \bigoplus \ldots \bigoplus H_N$$

- Where S is the quantum relative entropy:

$$S(\sigma_1\ ||\sigma_2) := Tr(\sigma_1 Log(\sigma_1)) - Tr(\sigma_1 \log(\sigma_2))$$

Appendix 2: QZE formal proof

- Evolution over time t of quantum system where $U(t) = e^{-iHt}$ is: $|\psi_t\rangle = U(t)|\psi_0\rangle$
- Survival" probability P_s that the system will still be in the initial state at t:
- From this we can derive: $P_s = |\langle\psi_0|\psi(t)\rangle|^2 = |\langle\psi_0|e^{-iHt}|\psi_0\rangle|^2$

$$P_s = 1 - (\Delta H)^2 t^2$$

- Where:

$$(\Delta H)^2 = \langle\psi_0|H^2|\psi_0\rangle^2 - (\langle\psi_0|H|\psi_0\rangle)^2$$

- This is an approx. valid for small t (expression becomes negative for $t > \Delta H$). Obtained by expanding e and only keeping the terms up to order t^2 in the survival probability expression : $e^{-iHt} = 1 - iHt - \frac{1}{2}H^2t^2 + \cdots$
- We may now define the Zeno time: $Z = \frac{1}{\Delta H}$
- so that: $P_s = 1 - \frac{t^2}{Z^2}$
- Now consider measurements separated by time intervals that become arbitrarily small. Replace single measurement after a time t with N successive measurements at time intervals $\delta t = t/N$. Survival probability becomes:

$$P_s^N = \left(1 - \frac{t^2}{N^2 Z^2}\right)^N$$

- So, as t tends to zero and N tends to infinity: P_s tends to 1. Strict m-property theory therefore cannot work.

Appendix 3: 1-particle master equations

- GRW:

$$\frac{d}{dt}\rho(t) = -\frac{i}{\hbar}[H, \rho(t)] - \lambda_{GRW}\left[1 - e^{-(x-y)^2/4r_C^2}\right]\langle x|\rho(t)|y\rangle$$

- KR:

$$\frac{d}{dt}\rho(t) = -\frac{i}{\hbar}[H, \rho(t)] - f\left[\varphi\left(\rho(t)\right)\right]\left(1 - e^{-(x-y)^2/\boldsymbol{g}[\varphi(\rho(t))]}\right)\langle x|\rho(t)|y\rangle$$

- CM:

$$\frac{d}{dt}\rho(t) = -\frac{i}{\hbar}[H, \rho(t)] - j\left[\varphi\left(\rho(t)\right)\right]\left(1 - e^{-(x-y)^2/\boldsymbol{g}[\varphi(\rho(t))]}\right)\langle x|\rho(t)|y\rangle$$