Dissolving Thought Experiments

(read about EC first, else you may be confused/misled)
(presented in no particular order)
(click links for descriptions, as only dissolutions are provided here)

(todo/note: I had to restore this document from a old unverified backup, so this may be incomplete/flawed)

Dissolving Thought Experiments

Chinese room

See also the section on Functionalism.

1. “Understanding” != consciousness, and neither does consciousness require it
2. Even if it somehow did, the mainline claim of this thought experiment - to say “understanding” can’t exist in non-biological or digital systems - is incorrect.

Those two points are very connected, since in investigating what we could mean by “understanding” - in Taboo’ing that word - we see it always functions as a reference to arbitrary capabilities. And according to either the hard problem or EC, we aren’t concerned with any particular capability, but with how anything - including any capability - could be conscious. According to EC, “Understanding” (given most reasonable definitions) isn’t required for consciousness (existence). But more importantly, for it to be required there must be some positive defense of this requirement, and how it connects with some definition/understanding of consciousness. Most (all?) people using this thought experiment have no such positive account of consciousness; to them it is still a mystery, and as such a requirement of this sort has no support.

On its own terms (not necessarily connected to consciousness), this thought experiment fails to demonstrate that non-biological systems can’t “understand” - because there is no absolute “understanding”. A Chinese Room, a GLUT, a human, an AGI – these are all doing different things under the hood. To answer if any of these things “understand” a given thing, we have to find a precise definition of understanding - to Taboo it - which can only ever be done with reference to a particular context.

What would it mean for a given AI agent in a video game to understand something? Perhaps in one context it's useful to say the AI “understands” the player character has item X when a dialog event is triggered, and then internally the AI adds X to a list of strings. We could then go on to explain the entire implementation stack of that particular “understanding”, from software to silicon atoms. We are not yet in a position to explain our full tech stack, but it is there, and it is different.

Similarly, building an argument based on a hard distinction between syntax vs semantics (Searle argues digital computers only have access to syntax - never to the semantics/understanding) is making the same mistake of not asking how we as humans can have a grasp of semantics. Some process is happening when a human “understands”. As a process, it yields to some computational description, and due to the concept of software - and the universality of computation - there is no reason that process cannot be instantiated/implemented in another substrate. This aligns with the anti-Zombie and pro-program arguments within the EC document.

Chinese brain

See the section on Functionalism.

Mary the color scientist / red Mary / Mary’s room

Taboo the word "know", which then reduces to a certain brain state. Thus <Mary studying the color> and <Mary experiencing color> end up in two different brain states. Her likely inability to shift her brain into the <experiencing state> only by sitting/thinking/studying doesn’t invalidate what else her understanding could accomplish. At a superficial level, through her understanding she would be able to manipulate her brain (with additional tech) to directly achieve the state of experiencing color. But more fundamentally, complete understanding would entail the ability to find any actual and hypothetical content within the software program. According to EC, by finding the content we will have found the consciousness of that content in full. (Aside: since Mary is in the black/white room, she isn’t “finding” it by scanning her brain or some such. The setup of this thought experiment is that she knows everything about the program, which would include the ability to extrapolate the consequences of hypothetical input.)

This thought experiment hinges on the claim that no matter how much we know about the brain or program, we will never know experience itself. It is an appeal to imagination, but a limited one that only begs the question. Something like “look at all those billiard balls bouncing around, look at whatever kind of complicated process/algorithm or chain of causality you imagine you’ll find in the brain - that doesn’t touch experience!” But in taking that call to imagination seriously, what do we imagine we will actually find with full understanding? The program and all of its content in full detail. Once we've embraced physicalism, the full fidelity experience is there to be recovered via some correct interpretation of the substrate.

This thought experiment's strategy - if applied to a conventional computer - would attempt to say something like "even once you have knowledge of the position of every atom in a laptop (excluding the monitor and prior knowledge of computers, to make this analogy work), and can predict every 0 and 1 and the timing of every circuit, you don't know whether msWord's UI is currently open on the desktop."

It is true that if all we are granted is the position of every atom, that alone strictly doesn't contain understanding of the program those atoms map to. That mapping, or layer of interpretation, is a kind of understanding that Mary - ex hypothesi - would have.

Absent qualia

Could be taken as an umbrella term for any proposal that says it's possible to have the same functional process, but for whatever reason there just is no consciousness/qualia. This is effectively an argument for zombies, which is already addressed within EC, which internally links to this dissolution.

Or, absent qualia could be taken to mean how someone like Dennett wants to do away with the term/concept "qualia" altogether - and/or even if temporarily accepting some definition, he argues that such a thing doesn't exist - for what I would regard as largely strategic/tactical reasons, which I disagree with in the section on Illusionism/Fictionalism. Also related to the section on the Brainstorm machine.

Absent qualia could also be confused with something like the visual blindspot or blindsight, both of which are addressed below.

Visual blindspot

The "invisible" gap in our vision from where the optic nerve passes through the retina. In connection to EC, can be used to demonstrate that absence of information != information of absence. There is no magical "filling-in" of our consciousness that needs to happen for us to be unaware of the blindspot. Related to the sections on visual saccades and the Marilyn Monroe wallpaper.

Visual saccades

Another demonstration that absence of information != information of absence. Consider the gaps in time between saccades - Dennett: "We don't notice these gaps, but they don't have to be filled in because we're designed not to notice them." Marvin Minsky: "Nothing can seem jerky except what is represented as jerky. Paradoxically, our sense of continuity comes from our marvelous insensitivity to most kinds of changes rather than from any genuine perceptiveness."

Related to the sections on the visual blindspot and the Marilyn Monroe wallpaper.

Marilyn Monroe wallpaper

Another demonstration that absence of information != information of absence. Also, how would we know whether we "had" "immediate" access to X amount of information, or whether we rather quickly retrieved it? Our conscious awareness flits all over our body/mind, much faster and with greater instability than most people realize. This is my favorite example to demonstrate that we can be very wrong about the nature of our experience, and what level of representation/detail is strictly required to reproduce EC-narrative-layer-style access to purported mental content.

Blindsight

A scotoma (blind spot) is present, but subjects have a better than chance (or perfect) ability to make guesses about certain events in the scotoma.

Any mystery invoked by this scenario is one of apparent engineering, not a deep mystery of consciousness. Take a hypothetical ereader program, which we stipulate will normally display the current page number in addition to the page itself. How would we think about a malfunction of this ereader, such that the page number was no longer displayed, but everything else continued functioning normally? There is still some current page open, that page still has some number/index, and that index can still be internally used and manipulated by the program. Only some subset of the ereader is no longer capable of “self-narrating” about some aspect of its operation. Similarly, blindsight subjects display varying levels of reaction to stimulus while reporting no consciousness of the stimulus. Is this so mysterious? There is some subset of the human brain’s program that is capable of such self report and narrative access (see the problem of access). With blindsight, there still has to be some neuronal representation that is used to provoke the unconscious reaction, but there is no universal reason why <the subset of the blindsight patient that is talking to you> should have access to it.

Hysterical blindness

That wiki/link isn’t great, so here is a passable definition from Dennett: “Sometimes people whose eyes and brains are apparently in working order, so far as physiologists can determine, have nevertheless complained that they have been struck blind; they support this complaint by acting “just like a blind person.” They are being sincere in their belief, but they unwittingly reveal there is at least some part of them that can see - via for example bumping into more chairs than even a blind person would - meaning some part of them would have to know where objects are, so as to to consistently run into them.

EC regards hysterical blindness as a derangement of the narrative-layer/subagent/whatever - wherein the map is further divorced from similarity to the territory, and some contents/representations are themselves in conflict with each other. All that can be described at an entirely functional level. There is no magical domain/subset/“finish-line” that privileges only the content that the narrative-layer has access to as the only content that is conscious.

Anton syndrome

The opposite of hysterical blindness; they are blind but they don't yet realize it. This has the same EC dissolution as in the hysterical blindness section.

Inverted qualia

Subsumed within the section on the brainstorm machine below.

Brainstorm machine

(related to Illusionism / Fictionalism)

In this scenario, our inability to tell what is the “right” orientation of the plug seems to imply that “color” is an example of a type of content that we actually wouldn’t be able to “find” or recover from interpreting the brain. This would be a problem for EC, which claims that there is no gap between the existence of recoverable (in principle) content and our consciousness of that content.

However, the thing the hard-questioner is attempting to find in this case - a particular color itself, independent of anything else - is a complete mirage. What we have are complex discriminators for color which consist of thousands of “facts” which are just details about the particularities of the construction and fidelity of that discriminator. Similar to the Marilyn Monroe wallpaper, the subject is mistaken about what representations are strictly necessary for them to have an experience.

Imagine the process of normalizing the brainstorm machine. There is an object -> seen by a person -> mental representation of the object extracted by the machine -> an attempted re-display of that content -> redisplay is seen by a person -> redisplay translated itself into mental content, which is then compared to the earlier mental representation of just the object.

At no point in this never-ending comparison + normalization loop (and its variants) will you ever find the color itself, because you are unknowingly trying to re-invoke an impossible step during the comparison. What is actually going on is you are unknowingly imagining/assuming that the color is redisplayed for some agent to view, leading to a circular loop of investigation for the chimera of "what the color really is", when really at any point in the loop you are just comparing representation vehicles and not the color itself. You can't step out of the normalization loop, or find some clever way to see the color itself, because there is no such extra thing. There are just the thousands of contextual "facts" about the discriminator and its discriminations/conclusions that we (the subset/narrative-layer) call colors.

Part of what makes color "ineffable" (ex: blueness of blue) is that our color discriminations are the "end" product of potentially thousands of facts about the incoming sense signals, the current state of our neuroanatomy, the contrasts between those signals (take visual illusions regarding color for example, which demonstrate that more is going on besides a simple translation between frequency of light to color, even if that is a major determining factor), etc. Like a jello box that has been torn in half, we have this highly complex discriminator, that ultimately just "reduces" what one can say about its discriminations to something like "its pink". If we want more information, the only things available are higher fidelity descriptions of all the factors/pieces of the discriminator.

There is nothing left subjectively for us to glean beyond that discrimination. That's what makes it ineffable, which isn't mysterious. How else could we experience color? There aren’t going to be little tags (in our experience!) like in a color-by-numbers book:

All we (the subset/narrative-layer) have access to are the “end” discriminations of color. It seems to us that there is more to color that is ineffable, but it also seems to us that there are actually hundreds of crisply detailed marilyns in our experience at any one moment, when there aren’t! The "blueness of blue" can suggest that there is an extra ineffable blueness that needs explaining but can't, OR we could just be restating the recognition that there is no such extra information at all, and thus any attempts to find some such extra information (that doesn't take the form of lower-level/highly-detailed information about the state of our various color discriminators - like the half of the jello box) will always fail, and thus the ineffability; the inability for our search to find something that doesn't exist. The blueness of blue is ONLY a restatement of the discrimination; there is nothing else. There's not enough room left for more explanation because it would have to fit within what we need to have access to, to reflect-upon and speak-about.

Hard-questioner: Why can't red be green? What would "fix" our experience of THAT color?

Me: What color? (we both understand I don't mean: "what english reference for what complex discriminator?" - rather, we both understand that you mean that qualia/experience of color which attends "red/green/etc")

Hard-questioner: Um…ah….THAT color!

Me: You see? The reference is circular; the color experience itself is ineffable in principle; we aren't appealing to coincident/contingent features of our inability to "access" some "deeper" understanding of color itself. Only reference to the complex discriminator/discrimination is possible. Even if we had reflective access to those thousands of "facts" about the discriminator, that wouldn't grant the type of insight the hard-questioner imagines is hidden. To reiterate at a more general level, it would violate physicalism for there to be some purely phenomenal change that we then could notice.

Color Phi phenomenon

(lost from backup; todo)

Cutaneous rabbit

(lost from backup; todo)

Zombies

Already addressed within EC, which internally links to this dissolution.

Twin earth - H20 vs XYZ

See also the section on Functionalism.
There have already been many takedowns of this thought experiment. But even if this scenario were possible, one’s beliefs about water - and the subcomponents of our substrate and program that would rely on the functioning of water - would be treating water as an interface which would be fulfilled by XYZ. In this case, the distinguishing details of XYZ vs H20 would be unimportant to the execution of the program. If we had a human brain upload running on a computer, we wouldn’t care what color the transistors were.

Addressing Traditional + Other Theories

Dualism:

Refer to premise 1 of the EC document.

Physicalism:

For a justification of physicalism as a whole, refer to premise 1 of the EC document.

Behaviorism:

By trying to explain consciousness via external behaviors and physiological measurement alone, you've prevented yourself from talking about consciousness itself. More concretely, we have many cases demonstrating this frame is false, such as locked-in syndrome.

Once we've escaped Behaviorism, things get madly confused and intermingled. There is a pattern of: theory X is proposed, which gets some things right and wrong -> theory Y is proposed as an antithesis of X based on a valid flaw of X, but then also throws away what X got right. It is a failure of nuanced synthesis.

Identity Theory:

Mind states are brain states. As a statement, this can possibly be interpreted as correct and in alignment with EC. However, other interpretations can produce fantastically confused and/or complex variations. This is also partially due to not strictly adhering to the specific and precise problem of consciousness alone, rather than the mind at large. I cannot go over every variation of identity theory, but one common potential complaint or flaw is that identity theory can too strictly identify with just the brain, whereas EC cares about the software program the brain implements. This leads us to functionalism.

Functionalism:

Mental states are constituted by the functional role and relation they have with respect to other mental states and the system as a whole, meaning mental states can be multiply realizable, meaning the mental state(s) can be considered as software that can be implemented via multiple alternative substrates. Again, there is a version/interpretation of this that aligns with EC. But if you consider the functional role of some component that is operating at a higher level of organization than the lowest isomorphic elements needed to accurately capture the program, then you miss those vital elements we actually care about - or in the extreme case you reproduce behaviorism by only caring about the input-output of the system as a whole.

Take the Chinese room: if you consider the algorithm/program the Chinese room describes, that program is different from the program being used by a human Chinese speaker, even if their output is the same. The internal details matter. We care about what is actually happening in reality in-full, but only to the extent necessary to capture our particular program. For example, we wouldn't care about the color of the transistors being used to run a brain upload, since even though that fact is in-reality, it plays no role in the algorithm/program.

Another species of criticism of functionalism is to reduce functionalism's claims to triviality or absurdity. If we are properly identifying the program we care about, then it could in principle be "implemented" by anything for which we could establish a mapping to that program. A blank piece of paper could in principle have a mapping to our program (assuming a large enough paper), which when instantiated would look like blueprints/write-ups of the program on that paper. Is the paper - either blank or with those blueprints - conscious? Or take the Chinese brain: would there be consciousness if billions of people each acted isomorphically to the behavior of our neurons?

The case of the Chinese brain is straightforward; EC would say that the human consciousness being emulated would actually exist, as the program is being actually run/executed, no matter how weird the substrate. Attempts to merely frame this as absurd aren't rational arguments. From the EC document:

"Consciousness is the quality/feature of there being "something that it is like to be something". In our case the thing "it is like to be" is the software program. If unconvincing, I would revisit the anti-Zombie argument. That article doesn't explicitly argue for this identification with software, but its reasoning can be reused here. If something is running the software program we care about, that then translates to: 1) that thing being capable (given sufficient peripherals/organs) of speaking/reflecting about its consciousness and all varieties of content in its internal experience, and 2) given we are identifying this software program, we are not identifying some cheap approximation of its output like a GLUT or a voice recording. So we would have an in-universe, physical explanation for why it is talking about consciousness the same way we do; it is running the same program. To then suppose that this thing wouldn't be conscious - only due to the hardware - would be to suppose the possibility of zombies; something that fulfills all the actually important properties of ourselves, but unlike ourselves isn't conscious. There is no reason to bite this bullet."

Similarly, arguments regarding this brain's supposed inability to interact or self-report deliberately miss that we only care about the program. If some human brain upload is running in an abandoned datacenter, that upload's inability to self-report is independent of the question of consciousness.

The case of the piece of paper is more complicated. According to EC alone, the paper isn't executing anything - even if the description of the program and all its possible states is complete. So by default, baseline EC claims the paper isn't implementing our consciousness. However there are broader metaphysical perspectives one could take, such that the hypothetical execution of the program is sufficient for that execution to occur in a platonic or alternative reality - which is just as real as our own. I will not attempt to defend this view here; I only provide this possibility for transparency. More mundanely, if one were to take that piece of paper and have someone execute it using yet more paper (follow the specifications/blueprints on the paper, provide inputs, trace state transitions, etc), baseline EC would claim that that implementation is just as valid as our normal brain.

To summarize: if you apply functionalism to a level of reality too high, you get behaviorism or identification with different programs which we don't care about. Applied to the correct level, the only available criticisms are unjustified appeals to incredulity. Applied too low, and you start considering details which aren't important to the execution of the program, which results in false inequalities (see also: Twin Earth).

Computationalism

I haven't read much about this. It seems to be scrambling in the right direction, but computation itself as a concept/process/dependency seems superfluous when you already have some concept of information/existence.

IIT (integrated information theory)

IIT's conditions for consciousness are arbitrary and ultimately baseless. It's not that hard to pick apart. I've heard this dissolution from Scott Aaronson is good, but I've yet to summon the motivation to read it (as such I don't endorse everything it may say). I don't mean to be too harsh on the originators of IIT; they were pushed by the same "impossible pressure" (mentioned in the main EC document) that forced crazy-sounding panpsychism to emerge.

Eliezer Yudkowsies' public thoughts

(another link/backup)

Eliezer's thoughts seem to revolve around aspects of the program (capabilities and contents) that he doubts are there for many (most?) animals. That is fine, but must not be confused with consciousness itself. We have to continuously re-insist and remind ourselves what we are referring to by "consciousness", as even notable hard-questioners use other meanings of that word - intentionally and unintentionally.

And really, it sounds like he is mostly invoking this "inner listener" / <self model>, which may just be another cartesian theater. But even ignoring the ways this might be defeated with anti-cartesian-theater arguments, the presence or absence of the self-model describe different things going on. But that there is anything at all is sufficient for EC. Is pain in X? Well, what is our contextual definition of pain? In a human we may recognize a type of pain by a certain pattern, which may be entirely absent in the pig. (but there are still surely some principals of mind-design that generalize)

He essentially seems to be operating with a definition/understanding of consciousness that is larger than used by EC. EC is very careful to focus very specifically on what it regards as a very narrow conceptual problem. If Eliezer is throwing too many capabilities into the mix - confused perhaps by consciousness historically injecting its tendrils of mystery everywhere - then it makes sense he reaches different conclusions.

However it's better to just focus your attention and definitions on the confusions/mystery; to get rid of this kind of deeply unassailable mystery and just have engineering left. Don't intermingle arbitrary capabilities because you will either never get the mystery out, or others will claim you aren't talking about the "consciousness" they are concerned with.

Illusionalism / Fictionalism

To insist that qualia and/or our consciousness don't exist is almost guaranteed to induce confusion and miscommunication, and in fact I believe Dennett had to start walking back "consciousness doesn't exist"-type statements for this reason. If using most people's direct understanding of consciousness - or at least the rough direction pointed to by the first section of the main EC document - I wouldn't say Dennett thinks THAT doesn't exist. To quote the EC document: "We can be entirely wrong about everything we have ever experienced - via for example living in a simulation, being a brain in a vat, hallucinating, or any other means - but the fact that there is experience itself is impossible to question."

My flavor of steelmanned hard-questioners don’t think that experiences have to be irreducible, or unexplainable, or that using qualia (or anything) for virtus dormitiva-esque explanations (ex: sleep caused by sleepiness, looking good in pictures because one is photogenic, etc) are valid, and they don’t have a settled opinion on whether experiences are/aren't the cause of anything.

Imagine all the sub-agents that are helping, in parallel, to create representations/content, deal with ambiguities and contradictions, model past/future content and behavior, etc - as all for the "purpose" of writing/fleshing-out this "fiction" of our heterophenomenological world, the subject's world. In this way does he call our phenomenology a fiction. Just like the conception/abstraction/fiction of a “center of gravity” in physics, there is a “narrative center of gravity". Dennett: "...and so a heterophenomenological text gets created. When it's interpreted, the benign illusion is created of there being an Author. This is sufficient to produce heterophenomenology."

He is trying to exorcize the phenomenal, but it is a misleading communicative mistake to call the entire thing a fiction or illusion, when really there are only aspects of the agent's intuitions that are actually fictions (its unification, how it feels itself to be the ultimate source of decisions, etc). When a subject says: "It seems there is a cup", 2 things are true (assuming they aren't lying/confused/misremembering/whatever): there is a seeming of a cup, but we don't yet know whether there is a reality to any actual cup being there. We are assigning reality to the seeming. That is our consciousness that the hard questioners care about. We want to explain that seeming - but not just the functional/physical way you can get a physical intelligent system/agent/robot to have those internal relations and produce such a narrative - BUT THE EXPERIENTIAL / PHENOMENAL / SUBJECTIVE / FELT aspect of having that seeming. If we were to discover/create some arbitrary bug in shakey's code that produced additional objects that aren't actually there under certain conditions - a visual illusion - shakey would reproduce a representation/report of that fictional object, and since we have a complete understanding of shakey we can completely explain - at a functional level - his misapprehension and subsequent behavior. What remains - and was our actual original concern - is whether THERE IS SOMETHING THAT IT IS LIKE TO BE SHAKEY. Whether he is conscious or not.

Another way of looking at the problem with calling our story a fiction: start by imagining a literal fiction: a Sherlock Holmes story, but in its physical manifestation; the physical book itself. The words on the page are indeed of a fictional person and world. Now what if instead, they described the actual/real physical book itself, its content now self-referential? What if this physical book could somehow walk/talk/interact? What if it gave itself a name? And the words in its "story" pertained to itself and its actions in the world? Now, not everything it thinks it knows about itself is true (its consciousness is less unified/uniform/continuous than it thinks, to take just one example), but at this point it is misleading to call its story a fiction. It is talking about itself - or its vessel, its physical embodiment as the singular body of the physical book - which is an actual thing in the real world. Its interpretations of its predicament, mental representations, and the external world may be flawed in any number of ways, but its story (our story) has to have some level of alignment with reality.

David Chalmers’ Topology

There is a very useful categorization of theories created by David Chalmers in his paper “Consciousness and its Place in Nature”. I only provide a summary of each type, as it should be clear by now that EC doesn’t strictly fit into any one; it is some transformation/combination of A,C, and F.

- Type A materialism

Fully material; no epistemic or ontological gap between the physical world and consciousness.

- Type B materialism

Epistemic but not ontological gap. Two different concepts may turn out to be equivalent in reality but we may have no way of conclusively reasoning towards that fact.

- Type C materialism

Epistemic but not ontological gap, but the epistemic gap can in-principle be closed with enough development of thought.

- Type D dualism

The physical world is not causally closed.

- Type E dualism

Epiphenomenalism; physical states cause phenomenal states, but not vice versa.

- Type F monism

Consciousness is constituted by the intrinsic properties of fundamental physical entities: that is, by the categorical bases of fundamental physical dispositions. On this view, phenomenal or proto-phenomenal properties are located at the fundamental level of physical reality, and in a certain sense, underlie physical reality itself.