Xinyuan Gu 顾馨媛 xinyuan_gu@outlook.com; keien@fuji.waseda.jp; 0212082@fudan.edu.cn

Revised in July 2017

From Color, to Consciousness, toward Strong AI

Abstract:
This article cohesively discusses three topics, namely color and its perception, the yet-to-be-solved hard problem of consciousness, and the theoretical possibility of strong AI. First, the article restores color back into the physical world by giving cross-species evidence. Secondly, the article proposes a dual-field with function *Q* hypothesis (DFFQ) which might explain the 'first-person point of view' and so the hard problem of consciousness. Finally, the article discusses what DFFQ might bring to artificial intelligence and how it might allow strong AI to stay true.

1. Introduction

Empirical evidence in cognitive science suggests that color is our common illusion of this world.

The nature of consciousness has long been under disputation. No consensus has ever been reached.

Can a machine think?

Color, consciousness, and artificial intelligence, three seemingly separated topics, are going to be discussed cohesively in this article. First, the article relocates color back into the physical world and reveals our largely shared, reliable reconstruction of it. Secondly, the article proposes a dual-field with function *Q* (DFFQ) hypothesis of consciousness, in which a hypothetical function could explain the hard problem of consciousness. Finally, the article considers some issues in artificial intelligence in the light of DFFQ.

2. Color

'Color' relates to how we see and how we understand the actual, physical world. Philosophical views on color seem to be chaotic. Among all the debates, questions such as 'Is "color" a property of objects?' and 'Is "color" mind-dependent?' serve critical roles.

This section of the article attempts to give answers to the questions. Here is a brief of color based on which the discussion will be developed:

We believe we see things in color, but scientific facts betray our intuitive belief. Physical objects invisibly 'interact' with lights; the latter are to be reflected, absorbed, and etc. Lights are not really passive at micro-level. Both the nature of objects and the nature of lights participate in determining the results of the interactions. We human beings believe we see objects in color, but we have good reasons to doubt if color is a part of the actual world or an illusion created and only perceived by our brain.

2.1 Function F and Function G

It is perhaps common for those who know the facts above to conclude that, since color is after all what we perceive, it is fine to say objects do not have color and to say seeing color is a kind of subjective experience.

Such a conclusion, however, is not satisfactory. One fact to be confirmed first is that, in theories and practices of physics, with known nature of objects and known nature of lights, all the reflections, absorptions, and other interactions, in principle, can be calculated. Let us represent the nature of an object with *x*, the nature of certain light with *y*, and the interactions in total with *z*. Knowing *x* and *y* should lead you to *z*; knowing *x* and *z* should lead you to *y*; and knowing *y* and *z* should lead you to *x*. Admittedly, if we only know Newton physics level details, it might be difficult to calculate quantum level results. However, the principle should hold. Due to this fact, the relation between the interactions in total and the nature of the two participators is *analogous* to a (mathematical) binary function: $z=F(x,y)$. Physicians have 'Parameter' as their term, but here in the analogy 'input' and 'output' are instead to serve the purpose of the discussion. In this binary function, *x* and *y* are the inputs, and *z* is the output. The calculations simulated by function *F* may involve a group of (mathematical) functions in actual practices; however, the relation is perfectly represented in $z=F(x,y)$, in the sense that the result of interactions in total depends on and *only* on the nature of the two inputs, namely the object and the light. Technically one would need all that involved in physical laws in a calculation. However, *x* and *y* are the only two actual-world dependent 'inputs'. Others involved are not actual-world dependent but law-dependent.

What does the fact say? It says that there is no place in $z=F(x,y)$ for perceivers. It says that, whatever the nature of the result is, the result does *not* depend on anything else world-dependent than the object and the light. It says that, without the presence of human beings, objects and lights will keep happily interacting with each other. On the other hand, without objects and lights, and especially without objects, human beings are not supposed to visually perceive anything, including color. Therefore, color objectivism is correct first of all on the point that 'the existence of colour instances……does not depend upon the existence of perceiving animals', as summarized and objected by Hardin (1984).

Why, then, would people so tend to emphasize the role of perception? *z*, as the output of function *F*, namely raw color, whatever it is, will be received by human eyes and further be processed by our brain. So 'The perceived colors are outputs of our brain' is a scientific description. Define color in our perception, namely seen color, as *i*, and the way a human brain processes *z* as function *G*, we will have $i=G(z)$. This function tells us that we are seeing a processed *z*, and how *z* looks depends on how we process it. However, remember *z* is the output of the *independent* function *F*. Therefore, function *G* could be rewritten into $i=G\{F(x,y)\}$. To those who over-emphasize perception, *z* is passive in and dependent on *G*; but the truth is, *z* is fully dependent on the result of *F*, and *F* is independent of *G*. To paraphrase, the color *we see* is dependent on human minds, but the *whatever* to be processed by our brain is *not* dependent on human minds.

2.2 Blue Things and Hue-Radiation Relation

As was pointed out in Hardin (*ibid.*), 1) 'Apart from their radiative result, there is nothing that blue things have in common' and 2) no hues are directly in or caused by lights (radiations).

Based on these two observations, Hardin concluded that 'objectivism fails'. This is the first challenge I would like to dissolve in this article.

Hardin's observations are probably correct; however, what do the two observations actually say?

In observation 1), all things that are blue have nothing (physical, chemical) in common. In other words, all *x* that cause *z*=blue in function *F* vary, which is fine in the *binary* function, because it is fine for no one-to-one relation between *x* and *z* in function *z=F(x,y)*. And in daily life examples *z* will vary within a range while still be taken as blue, so it is perfectly fine for all *x* that cause *z*≈blue to vary in function *F*.

Observation 2) is based on hue, saturation, and brightness. Define hue as *a*, saturation as *b*, and brightness as *c*. Color is a *z* that can be precisely represented in a combination of certain *a*, certain *b* and certain *c*. Light, as I defined, is *y*. So observation 2) says that no *a* is directly in or caused by *y*, which is still perfectly fine in the *binary* function *z=F(x,y)*.

The analogy, function *z=F(x,y)*, demonstrates a possible explanation for the two observations; therefore, Hardin's conclusion that 'objectivism fails' fails.

2.3 The Whatever to be Processed

Between the output of function *F* and the output of function *G*, namely, between *z* (raw color) and *i* (seen color), there is a gap. The gap is disturbing because if it is left unfilled, it will prevent us from asserting that we do know the actual world.

So what is *z*, namely raw color as the output of function *F*, namely a result of the interactions between an object and lights, namely the *whatever* to be processed as the input in function *G* by our brain?

A subjectivist view on processing *z* was refuted in Hardin (*ibid.*) based on the reason that it is almost impossible to specify 'normal conditions' for seeing (raw) colors.

However, our perception of color is not as random as described. It is beyond doubt that in various societies, practices based on perceived color are basically smooth. The only reasonable assumption is that we have a brain module which allows us to largely share perceptions on color.

Further, this article proposes that *i* is a reliable ***re***construction of *z*. To paraphrase, the color we see reliably reconstruct the raw color of the actual world for us.

Paintings are to be our hero. Suppose we have *z*1 as raw color in total of a beautiful scenery in front of our own eyes. A painter beside us just painted a realism painting of the scenery. The painter processed *z*1 with her eyes and then turned seen color *i*1 via her practice to raw color *z*2 of the painting. Now there is a forger beside us copying the original painting. The forger processed *z*2 with his eyes and then turned seen color *i*2 via his practice to raw color *z*3 of the copied painting.

Now we have the scenery in raw color *z*1, the original painting in raw color *z*2, and the copied painting in raw color *z*3 in front of us. And we are going to process *z*1, *z*2, and *z*3 with our eyes. The question is, are we going to say *i*1≈*i*2≈*i*3? And if the answer is yes, *what* makes that possible?

Moreover, we have chameleon (Chamaeleonidae) as a color ally. Think about this: we put a red object close to a pet chameleon and it would 'align' the color of its body to the one of the object. We saw the process with our own eyes and we believe that the chameleon is now in

the same color as the environment is. Obviously, the probably-inbuilt mechanism for chameleons to do color changing *agrees with* our function *G* upon environmental color. A chameleon may also be in a protective color, and if *to other animals* the color of its body is not aligned to the environmental one, there would be a much higher chance for it to be eaten. *What* makes that possible?

Should we trust ourselves *and* chameleons? Despite a chameleon might not see our seen color red as red, the environmental color is certainly beyond our skin and beyond chameleons' skin. Moreover, our function *G* and a color-dealing function of chameleons both reliably reconstruct the environmental color so that we and chameleons can reach an agreement. We may further generalize that protective coloring confirms raw color in the actual world.

Therefore, it is reasonable to define *z* the raw color as an extrinsic property of objects because *z* is a non-random status of appearance for *x* to be in when under conditional *y*. That is to say, color is a determined status of appearance for an object to present based on the physical-chemical nature of the object when under a conditional light. It is extrinsic because it is related to the external world.

2.4 One Missing Piece

Cognitive science says the number of cones employed by species differs, so some species detect more kinds of lights in various wavelengths, and some detect less——so the actual world in the eyes of different species would differ in its appearance. This is not a problem. Color could be like distance (alternatively, gravity). Suppose there are two rocks spaced out one meter apart. We human beings reconstruct the distance via visual perception so that we can avoid both rocks when walking; bats reconstruct the distance via sonar perception so that they can stand on one of the rock and precisely fly to the other. Our visual perception and bats' sonar perception, though not exchangeable, both reliably reconstruct the physical world beyond ours and bats' skin. Such a guess would assign *interobjectivity* to color.

But there is still one missing piece left. Color has to be an appearance of objects beyond our skin. But why 'red'? Newton proved that, sunlight can be divided into lights in seven colors generally visible to human beings. Why *the* seven colors? Since the reconstruction realized by our brain is to be trusted, Newton's red light of certain wavelength is not only seen as red but also *in a status* that can be perceived as being in red to us. What is *the* redness? Is it just a 'value' assigned to the certain wavelength based on human beings' biological foundation that helps us to understand the actual world, or something else? After all, our brain needs a starting point for sophisticated processing: either redness is physically-chemically determined by interactions between the brain and its environment, or it is a random 'value' our brain chose or happened to develop in its remote past. It is very unlikely that our ancestors were in an environment in which they frequently encountered the *clear-cut* seven colors in the Newton experiment. Despite so, we can perceive the *clear-cut* seven colors. This would be the remaining fundamental challenge to theories regarding the nature of color.

Although the puzzle remains, I believe it is appropriate to conclude here that color is an extrinsic property of objects, that we largely share the perception of color, and that the perception probably contributes to a reliable reconstruction of the actual world.

3. Consciousness

This section of the article proposes a **dual-field with function *Q* hypothesis of consciousness (DFFQ)**. The hypothesis aims to provide a primarily reasoning-based explanation compatible with some of the most important questions raised in the study of consciousness.

The word 'consciousness' is notoriously confusing. The 'Consciousness' entry in *Stanford Encyclopedia of Philosophy* lists concepts of consciousness discussed in numerous literature, including sentience, wakefulness, what it is like, and various conscious mental states. It is likely that consciousness is a galaxy having many moving stars obeying certain rules. We started with calling the galaxy 'star' and during our trip of discovery, due to the overwhelming richness of the reality and the limitation of the existing word, have to introduce more and more descriptions. When a guess on the galaxy is made, the meaning of 'consciousness' would be clear.

DFFQ assumes that an actual world exists. A part of living things are capable of *treating* the world as an information package. Since the world is *to be seen* as an information package, no world, no information package. In the previous section, color as a kind of information has been returned into the world. Consciousness is to be considered as a tool kit for a creature like a human being to deal with such information. And functions are rules followed in the tool kit. Computational terms are employed in this article only because analogies of computer would improve our understanding.

Regarding to eliminate or to keep the Cartesian theater, this article takes a position of considering the Cartesian theater as a wrong-in-nature explanation but deserves more than cancellation. A cursing fact is that we cannot call physical pain back just by thinking of it, yet we can recall felt-psycho-painfulness just by thinking of it. Similarly, we cannot call a red object back in front of us just by thinking of it, yet we can recall the, well, 'mental image' of the same red object just by thinking of it. In order to explain this, I have to disagree with the claim saying 'no such thing as an after-image or a sense-datum' (Smart, 1959). An after-image could be a piece of physical-chemically coded information as a useless byproduct generated by our brain.

By the end of this section, a coherent guess will be made on the Cartesian theater and on the hard problem of consciousness (for the latter, see Chalmers, 2015). The hypothesis will also support extending 'access consciousness' (Block, 1995b) to information other than visual. Moreover, comments will be made on Schacter's Model and its modifications (introduced in Block, 1995) and on Information integration theory (IIT) to further clarify DFFQ.

3.1 Function Q

In the previous section discussing color, the word 'function' has been used multiple times. It not only refers to mathematical ones but also represents specific, settled ways of handling certain actual world information. Realizing a function, such as realizing the function of seeing with human eyes, means extracting and processing certain actual world information in a certain way.

It is unlikely that consciousness involves one and no more kind of information. 'The hurtfulness of pains, the itchiness of itches, pangs of jealousy, … the characteristic experience of tasting a lemon, smelling a rose, hearing a loud noise or seeing the sky' (Jackson, 1982)

described as qualia are all based on information of different physical natures.

Consider this: Nowadays a smart TV allows us to replay recorded contents. Sometimes audio goes slower than video and your brain struggles to synchronize them. Consequently, it is hard for you to *understand* what is going on. When we talk about human eyes, we are comfortable to say that rod and cone cells play different roles in covering visual information; but are we comfortable to say that ears and eyes play different roles in understanding? Auditory and visual information are physically different and are handled by the function of hearing and the function of seeing realized by organs, respectively. Understanding seems to be another function integrating different kinds of information. The failure of understanding suggests the importance of synchronization.

Further consider synchronization with this:

We may say 'She picked up apple and banana strictly simultaneously'. The sentence is odd yet tolerable as the subject may accomplish the verb 'pick up' strictly simultaneously with two hands. But can we *feel like distributing our attention on* seeing something particular and smelling something particular strictly simultaneously?

The distinction between phenomenal consciousness (P-consciousness) and access consciousness (A-consciousness) proposed (Block, 1995b) could be extended to explain synchronization and integration. A-consciousness is supposed to have access to every kind of experience described as qualia and further thoughts, desires and emotions. For instance, I am enjoying a painting in a museum. My eyes are collecting necessary visual information from the painting. My *access* to the visual information of the painting as a whole makes other sensible information, such as the temperature in the museum, a scent of someone, a talk in low voices, fade into the background. My skin, nose, and ears are still doing their jobs. Only that their works are not recognized at the moment by my A-consciousness. And at first I do not know where to look, so I randomly direct my attention on this and that details of the painting. This is to say I am accessing some particular visual information and making other visual details fade into the background. Now a professor of semiology joins me. She starts to point here and there of the painting and explain them to me. Suddenly all the details look different as a layer of interpretation has been added. And adding a layer of interpretation likely requires the function of understanding.

We can even extend A-consciousness to discussions in the wider context of biology. We train dogs to search for explosives for us. In such trainings, we would want a dog to remember something smells like *X* and have the dog to search for *X* in an area. Note a dog has the function of smelling all the time and all kinds of smells are always *accessible* to it. For a dog to complete the task of searching for *X*, it would have to put *X* at the center of its A-consciousness and make other smells fade into the background.

Remember there is a fair demand for an explanation of 'the first-person point of view', 'subjective aspect' and 'internal access' (Chalmers, 2015). To explain those is to explain the hard problem of consciousness. Along with all the above discussed, what I propose is a hypothesized function *Q*.

'Single *Q* at a time (synchronic)' accompanied with 'single *Q* and always this *Q* (diachronic)' would perfectly explain 'the first-person point of view' and the 'subjective aspect' and would be compatible with qualia and A-consciousness. Define heard sound=*a*, what we see=*b*, pain=*c*, emotion=*d*…, and put these information provided by ears, eyes, skin,

and brain itself into function *Q*, we will have *Q*(*a*), *Q*(*b*), *Q*(*c*), *Q*(*d*)…. Inputs *a*, *b*, *c*, *d*…are to be reviewed by function *Q*. Put, for instance, pain=*c* into *Q*, we have *Q*(*c*). What is the output? I am examining my pain, and what I gained is my experience of feeling pain, that is, painfulness. And we have discussed that our call-back-bilities upon pain and painfulness are different, so pain and painfulness cannot be the same. The puzzle of transparency (Tye, 2015) canceled the Cartesian theater by denying phenomenal character with a price of denying experience. An alternative explanation with function *Q* says that we are examining an *immediate output* of seeing, and there is *nothing else* to be examined, which also explains the transparency. Function *Q* registers what has been reviewed as experience and the reviewing history established is our memorized experience that can be recalled. 'Internal access' is a description of *Q*'s accessibility to various outputs. The so-said privileged access is limited: we can review our thoughts and feelings but *never* detailed processes of our brain (and body) functions. For instance, we do not access to the internal organizing process of our language.

This would also explain why we do not feel like distributing our attention on seeing something particular and smelling something particular strictly simultaneously. Just as the function of seeing, function *Q* might process only one input at a time. These two conversations may illustrate what does 'one input at a time' mean to a single function:

A: I ate chicken and peanuts strictly simultaneously for lunch.
B: You ate Kung Pao chicken? (chicken and peanuts summarized into Kung Pao chicken)

A: I ate Tiramisu and hamburger strictly simultaneously for lunch.
B: …….Why did you do that? (cannot integrate Tiramisu and hamburger)

So if two pieces of information can be integrated into one, then there would leave only one piece of information to be processed. The puzzle of bistable image would thus be explained: we know that people can only process one interpretation of the bistable images at a time. A bistable image is one piece of visual information, which when taken as an integrated one would mean nothing to people. One interpretation based on the image is Interpretation-1, and the other interpretation based on completely the same image is Interpretation-2. Since one cannot integrate Interpretation-1 and Interpretation-2 into one meaningful interpretation, one would tend to switch between two meaningful interpretations based on the same image. For the image to be processed, some function, most likely understanding, is necessary, and as the function cannot process two separated inputs at the same time, function *Q* accessing can never obtain Interpretation-1 and Interpretation-2 at the same time. Though we have not yet clarified the relation between the function of understanding and function *Q*, we might also guess that the smart TV synchronization problem is due to the unfulfilled condition of 'one (integrated) input at a time'.

To summarize, *Q* is such a function realized by our brain:

1) Single *Q* at a time;
2) Single *Q* and always this *Q*;
3) Single *Q* processing one input at a time;
4) Single *Q* that has selective access.

3.2 Distinct Understanding from Q

Empirical evidence to some extent support such a theory: that understanding is a separated function from function *Q*.

The most dramatic cases are introduced in *Who's in Charge? Free Will and the Science of the Brain* (Gazzaniga, 2011:Ch.2&Ch.3). The first study says a split-brain teenage patient, after heard 'who is your favorite' with both hemispheres and the continuing part 'girlfriend' only with the right-hemisphere, had emotional responses due to the word 'girlfriend'; the second study says a split-brain female patient, when encountered a picture of 'pinup girl' only with the right-hemisphere, she snickered. In both cases, both patients' left-hemispheres failed to correctly report what had happened. The two cases are somehow different from flashed-words experiments. In the latter ones, normal people would not notice the fact of being flashed by certain words, but would make decisions based on those flashed information. In flashed-words experiments, the sensible visual information certainly reached somewhere without being registered by function *Q*. One might argue that responses based on flashed words suggest a network connecting words and morphemes, but this is not the case in the split-brain ones. First of all, the verbal information 'girlfriend' and the visual information 'pinup girl' are both more or less not as 'timeless' as items such as 'snake' is, which means to respond like the patients did, one would need not only a semantics-syntax based but also a context-based understanding. And the *word* 'snake' would also be different from a picture of a snake. It seems that, the teenage patient who heard 'who is your favorite' with both hemispheres and the continuing part 'girlfriend' only with the right-hemisphere must have combined the two parts into one sentence with his right-hemisphere, and laughed based on the contextual meaning of the entire sentence. So the responses given by the patients are not induced by inbuilt emotion triggers but seem to be caused by a thorough understanding of the given information, which is odd as the left-hemisphere is thought to be responsible of dealing with language.

And in both cases, the oral reports from the patients are worth noticing. The 'girlfriend' boy didn't know why he was laughing. The 'pinup girl' woman said she saw nothing and faked a reason for her laugh. The oral reports are coherent with other faked-reasons experiments introduced in the same book (see *ibid.*). This also reminds us with the blindsight cases and the anosognosia case examined by Block (1995b). In the blindsight cases, patients with their primary visual cortex damaged were not supposed to be able to see anything, and they did report that they didn't see anything, and yet some of them managed to make decisions possible only when they did 'see' something. In the anosognosia case, the patient did not take her problem of recognizing faces as a problem and did not take her unusual efforts as unusual. One more case that might be related is the one introduced by Mason in her lecture (Understanding the Brain: Neurobiology in everyday life, Week 3, Language circuits), saying a born-blind patient had her visual cortex damaged by a stroke. Thereafter she could orally report that a bump is a bump, but could not tell the meaning of the Braille words anymore.

Another case regarding visual reconstruction also introduced by Mason (*ibid.*, Week.4, Learning to see) might help us to integrate all that we have explored. The patient had his cornea completely damaged at the age of three and a few months. When he became an adult, he had a corneal transplant, which means he then had full-functioning eyes. He thereafter can deal with motion and color (with which he could deal before the damage had been made), but

he has problems in recognizing faces and in telling implications of similar visual items.

Taking visual information as a place to start, a theory compatible with everything above might be like this: those human beings who can perform the function of eyes extract certain visual information from the actual world. The extracted visual information would be reviewed and registered by function *Q* and becomes a part of our experience. Understanding is an independent automated function that works on such as visual information *not necessarily experienced by function Q*. As demonstrated in Figure 1:

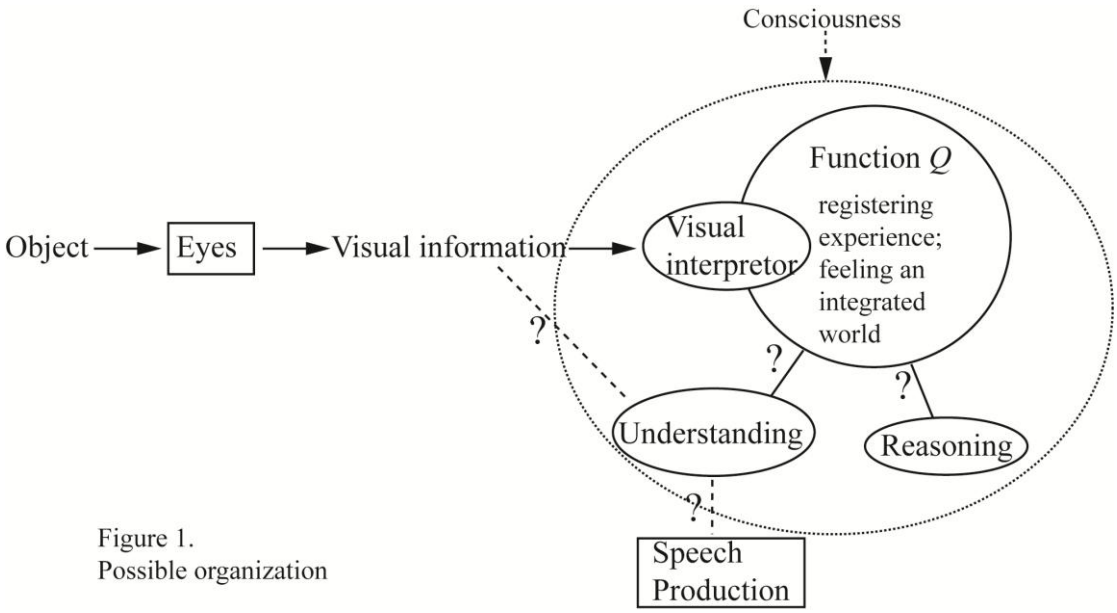


Figure 1.
Possible organization

Five cases, namely the woman seeing pinup girl picture, blindsight, anosognosia, the meaning of Braille, and the corneal transplant, all involve a chain or pathway of see, aware, and report (here 'aware' is used roughly as a synonym to 'be conscious of').

The corneal transplant patient has his eyes as a normally functioning organ, which means what can be extracted by eyes would be correctly extracted and passed to the brain. He can report that he has such and such problems, and the reports fit *what we observe*. So we may infer that the 'aware' (namely function *Q*) and the 'report' (speech production) functions are fine. The impairment seems to be about one or more particular functions that are supposed to add an interpretation to the information given by eyes.

The born-blind patient's visual cortex was compensating to deal with Braille. We may say she had her function of seeing compensated by having somatosensory inputs passing her visual cortex. In other words, the 'see' part is to be taken as fine at first. After the stroke, since she correctly reported that a bump is a bump, function *Q* and speech production both did their jobs. What has been lost is the layer of meaning added to those Braille words as pure somatosensory information given by hands. The correct link between the bump in the actual world as an object and the 'bump' as a word or signifier shown in her oral report, might be closer to the flashed-words experiments, suggesting a separation of networked words(if such a network exists) and the interpretation of words.

The blindsight cases are complicated. Whether the patients saw the objects or not was at first unknown. It could be that, the patients' function *Q* did not have the experience of seeing

(or, A-consciousness didn't have access to the visual information); oral reports governed by function *Q* said seeing nothing; oral reports led by information not experienced by function *Q* made correct guesses on the visual information extracted from the objects; motions such as 'grasp' correctly controlled the objects——which means the patients to some extent must have their function of seeing worked, and means catching the features of the objects does not need the participation of function *Q*, and means the function(s) that correctly caught the features of the objects might have a direct communication with speech production and motion without receiving the supervision from function *Q*.

The split brain pinup girl picture case and the anosognosia case are somehow most alike, as in both cases we know the patients had a fine function of seeing, and we know that their oral reports did not fit what we know, but we do not know if there were particular mistakes made by their function *Q* governing the oral reports, or if there were some particular functions had direct communications with their speech production (which implies that function *Q* only registers the outputs from such particular functions but not governs them).

At least it is likely possible for understanding to work independently upon the information provided by functions such as seeing. This would suggest an even complicated guess, comparing to Schacter's model and its modifications, on how might 'consciousness' work.

3.3 The DFFQ Hypothesis of Consciousness

Schacter's model and its modifications discussed by Block (1995b) include response systems, specialized modules, phenomenal consciousness, executive system, procedural/habit system, and declarative/episodic memory. To avoid confusion that might be brought by any further modification, what I propose instead is a **dual-field with function *Q* hypothesis of consciousness (DFFQ)**.

DFFQ needs three components, namely **Basic Functions (BFs)**, **Coordination Functions (CFs)**, and **function *Q***. It claims that our brain and body might work in a dual-fielded processing system.

BFs automatically extract information directly from the actual world, always give out information, and could be called into specific tasks (e.g. a dog searches for *X*). BFs may not call tasks themselves, but may call attention and responses by reaching various thresholds. The function of seeing, hearing, smelling, motion, (possibly) operation of semantics-syntax, and others are all BFs.

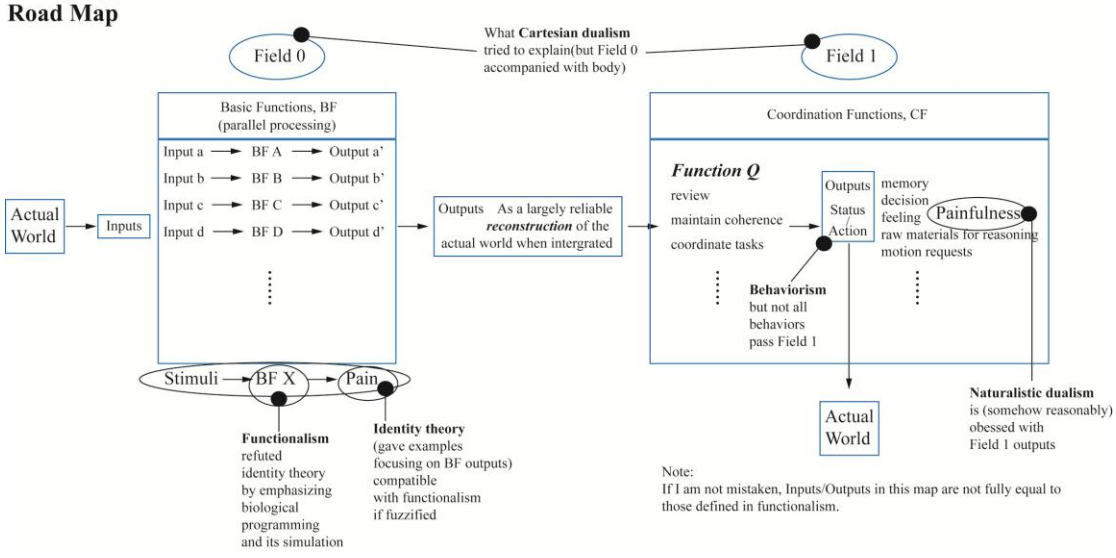
CFs receive information from BFs, reconstruct the actual world and interact with it, give out further processed information, may call tasks independently. CFs include understanding, reasoning, memory, and others.

Function *Q* is the not necessarily special one that selectively reviews the information given by BFs and by CFs (such as thoughts, emotions), registers information as experience, enables the so-felt first-person point of view(or, as the first-person point of view itself), and constructs a coherent reviewing history for its inputs. Function *Q* makes a coherent 'self' possible, but itself is not equal to 'self'; as 'self' is a history generated, not the generator itself. The experience for us to feel like distributing our attention might be due to the limited processing resource (one input at a time, like other functions) function *Q* has. 'Attention' appears when function *Q* focuses on reviewing specific information, on registering specific experience, and on calling specific tasks.

Some CFs might always be called by function *Q*, while other CFs such as understanding might work independently. One definition of consciousness describes consciousness as 'awareness or sentience that begin when one wakes in the morning and continue throughout the period that one is awake' (Searle, 1990b, cited in Block 1995b). In the discussion of the corneal transplant case, we have mentioned there might be one or more particular functions that are supposed to add an interpretation to the information given by eyes——and if such functions exist, they are more likely to be highly self-governed. As long as our functions of seeing, hearing, smelling and other BFs work well, when we are awake (or, when function *Q* is working), we cannot 'shut down' the world sensed, unless we close our eyes, cover our ears and nose, or retreat our body from everything. Even when we, for instance, close our eyes, as long as we have an activated function *Q*, the reconstructed world is still there. This explains why we would be managed to search for something behind our back with our hands without turning our head when we know that something must be behind our back: function *Q* might also be responsible for keeping registering a coherent real-time history of the actual world. But the reconstruction is probably done by some CFs that cannot be 'shut down' when function *Q* is working. Take one's wife as a hat, cannot feel one's own healthy leg, fail to recognize faces——these symptoms are likely about certain missing pieces caused by certain damaged functions that are supposed to participate in the automated reconstruction of the actual world. On the other hand, CFs such as reasoning are unlikely to be always 'on', even when function *Q* is at work.

Define the field where function *Q* works along with CFs as **field 1**, and the field where BFs work as **field 0**. Such a road map is demonstrated in figure 2:

Figure 2.
DFFQ Hypothesis of Consciousness Road Map



DFFQ says that ***consciousness*** is an **operating system (OS)**, or, an environment, coordinated by function Q, in which various field-1 input-to-output processing are possible and some accessed/called. It is possible for consciousness as an OS to be fully 'reduced' to the brain processes involved, so in a sense consciousness is *identical* to a certain organization and functioning of brain processes.

Roughly speaking, BFs include the response systems and the specialized modules in Schacter's model, while CFs include the phenomenal consciousness and the executive system. DFFQ resolves the procedural/habit system and the declarative/episodic memory in Schacter's model. The former is considered as some kind of independent know-how CF calling BFs such as motion into work following certain records based on learning, which would explain why when one is unconscious some habitual motions could still be well performed. The declarative/episodic memory in Schacter's model is considered as certain CFs accepting registrations from function *Q*.

In summary, our brain and body might generally maintain a single-'subject' consciousness as an OS, work in a dual-fielded system, have threshold-reaching and task-calling as two kinds of conversation initiating ways, and have all those functions to handle the actual world. The outputs of the entire DFFQ system would participate in the actual world. And the altered world would again become an altered information package to be extracted.

Cartesian dualism might have tried to understand the system by wrongly dividing body and mind into different substances; naturalistic dualism might have focused on field 1 outputs and wrongly believes those outputs are 'over and above the physical'; functionalism seems to be larger than the DFFQ map, but I would like to highlight (if I am not mistaken) its emphasis on functions in its original sense here; identity theory was taking field 0 outputs as examples and canceled field 1 ones.

Interestingly, functionalism and identity theory may actually have focused on two different parts in a chain of processing. This is where I feel 'Pain=C-fiber firing' could be fuzzified. 'Pain=C-fiber firing' in Smart (1959) and Place (1956) would mean 'consciousness=fibers firing', which makes consciousness nothing else but a brain process. There probably was a presumption in identity theory that the discussions were about human consciousness, thus when Smart and Place gave 'Pain=C-fiber firing', they actually meant '(human's experience of) pain-C= (human's) C-fiber firing' (though technically they canceled field 1 outputs). It looks perfectly fine to further have '(ape's experience of) pain-D= (ape's) D-fiber firing'. And in general to have '(a conscious living's experience of) pain-X= (a conscious living's) X-fiber firing'. Identity theory requires type-type identities, but is pain-C identical to pain-D and to pain-X (note the word 'identical' has a different coverage)? It looks perfectly fine for pain-C to share the *broad-type name* '*pain*' with pain-D, just as a human's eye can share the *broad-type name* '*eye*' with a fly's compound eye. In other words, despite that pain-C may not be 100% identical to pain-D due to the difference in the physical-chemical/functional aspects of C-fiber and D-fiber, pain-C and pain-D are both pains that can be experienced. It would be, however, biologically impossible for a rock to have pain and to experience it. Different OSs of different species in this sense certainly share the broad-type name 'OS'. And despite overcoming the experiential isolation between human beings and bats (Nagel, 1974) is still difficult or impossible, at least we may say some outputs such as pain are broadly shared among species while some are not.

3.4 Bridging Consciousness and Artificial Intelligence

This part of the article covers two topics, one is the principle of realizing an artificial consciousness, and the other is clarifying the meaning of integration different from the one in the Information Integration Theory (IIT).

First, regarding the principle of realizing an artificial consciousness. To various species, there is one physical world on this earth as an information package (IP) to be handled. Suppose there are only two species having consciousness as OSs encountering the physical world, the world is analogous to a zipped computer file that contains many sub-files. Species K and species U are to extract sufficient files *following specific rules*. Although it is appropriate to say that DPK and DPU are mind-dependent, the *rules* that generate DPK and DPU are *not* mind-dependent. K would generate a domestic package k (DPK), while U would generate a domestic package u (DPU). A zip file containing a '.dmg' file (mac OS executable format) could perfectly be present when the zip file is unzipped in Windows, but the '.dmg' file would remain inaccessible as there is *no rule* in Windows designed to read the '.dmg' file. Similarly, there would always be something in IP (e.g. particle movements) inaccessible to the biological foundations of K and U. Instead of keeping the inaccessible parts domestically, DPK and DPU 'abandon' obtaining the inaccessible-s. If K is a human being, then it is impossible for DPK to contain sonar files. So, DPK and DPU are necessarily different, both in rules and packages generated. However, the non-random rules allow DPK and DPU to contain reliable reconstructions of IP. DPK and DPU following different rules establish reliable reconstructions of IP, is a theoretical bridge available to fill the gap between the physical world and the perceived. A human being would run into a glass door, and a cat would, too. The glass door belongs to the physical world independent of the human's mind and the cat's mind, despite the domestic packages established are different. A robot could perfectly avoid the glass door if it is equipped with certain sensory components, not because the glass door *to it* exists but because it follows a rule sensing and telling that there is an obstacle to be avoided. So establishing an artificial consciousness is to establish an OS reasonably structured that can efficiently and largely reliably deal with the actual world. Such an OS does not have to be the same in every aspect to ours.

Synchronization and integration could play important roles in establishing an artificial consciousness. But the meaning of integration may need some clarification. The Information Integration Theory (IIT) proposes an 'integration of sufficient quantities of information', according to some articles at the website 'Conscious Entities'. In DFFQ, by contrast, integration could be done upon information of *different natures*. Rather than doing evaluations based on the amount of information accumulated, integration in DFFQ does a further process on known different information and outputs one piece of information of yet another nature.

An analogy can be made this way: there is a rock band, in which there are four members. Vocalist John, guitarist Tina, keyboard Li, and bass Eric. Each member is seen as a single core. Four members make a physical quad core. Subject John (SJ) is capable of writing lyrics (SJV1, V=verb) and vocal (SJV2); subject Tina (ST) is capable of composing (STV1) and playing guitar (STV2); subject Li (SL) is capable of bakery (SLV1) and playing keyboard (SLV2); subject Eric (SE) is capable of doing choir (SEV1) and playing bass (SEV2). The verbs are called 'threads'. Each core has two possible threads, so the physical quad core has eight possible threads as a community and five threads when performing (SJV2+STV2+SLV2+SEV1+SEV2). Our bass Eric in performance has to distribute his attention to two threads. If he was a computer, one of the threads would be virtual. A performance is something constituted with lyrics and music but of another nature as a whole,

just like Kung Pao chicken constituted with chicken and peanuts (but recalling the discussion on the 'concept' of mind (Ryle, 1949), note a removal or change in lyrics or music would make a performance at least to some extent different). The integration seems to be done by the function of understanding upon auditory information and visual information is another example.

In this sense, integration is a process, in which information of nature x obtained by function X and information of nature y obtained by function Y to be integrated by a third party function Z generating information of nature z, and it is certainly fine for information x, y, and z to quantitatively accumulate. The integration proposed by DFFQ is a function-nature based processing of information. For our brain, rules are functions. Failures in facial recognition might imply a damage of functions or a damage of a database on which functions are functioning. DFFQ also guesses that there can be a function-module dynamics ('module' is in the evolutionary psychology sense): a single function could realize a single module; a group of functions could realize a single module; and a re-group of functions could realize a (e.g. compensating) module after appropriate trainings.

4. Strong AI

4.1 The Problem of Understanding

The Chinese Room argument (Searle 1980; 1990a), receiving numerous comments, is a famous challenge to strong AI. The argument states that, suppose there is a room driven by a 'purely formal' computer program translating English language inputs (which are just symbols) into Chinese language outputs (which are just symbols as well); since the program can pass the Turing test without *understanding* the '*meaning*' of the inputs and of the outputs, strong AI is false.

Despite the conclusion strong AI is false is arguable, the Chinese Room itself rather has demonstrated that performance would not guarantee everything strong AI wants. If strong AI claims that an 'appropriately programmed computer' (Minds and Machines, slide 2:6) has 'mental states', to say that the Chinese Room does not falsify strong AI, one option is to prove that the Chinese Room is not 'appropriately programmed'.

What has been missing? As stated, understanding. The Chinese Room handles a difficult task, namely translation, with a very simple structure:

> symbols of English language as inputs → an artificial module of translation → symbols of Chinese language as outputs

The artificial module translation might have a grammar thread and a symbol-converting thread, but the Chinese Room is still single-fielded. Could we say that it is impossible for a single-fielded structure to simulate understanding, or thinking, or consciousness in general? Might we create field 1 and add an artificial function of understanding to the Room? Might we further create an artificial function *Q* to keep the function understanding on the track and review every input and output of the Room? If a revised Chinese Room has an artificial function *Q*, an artificial function of understanding, and the translation module at work, might we say the Room is literally *thinking*? Does it literally have mental states?

Simulating the hypothesized independent function of understanding could, however, be the

most difficult part; as the function has to do with world-relating information, and the understood information has to cause what one *wants* to express in another language, otherwise, symbols are just symbols and nothing is counted understood.

A make-do strategy could be like this: suppose there is a dictionary Mary imprisoned in a room (find the original black-and-white Mary in Jackson, 1982). She learns everything about the actual world through dictionaries and pictures, which means she only has second-hand information about the actual world. Is her stored information world-relating? To make the question simpler, suppose a child once saw a picture of cat with the caption 'cat'. When the child encounters a real cat, he would be able to recognize it (one interesting thing is a human child could do it without viewing hundreds of cat pictures). The question is, is the stored information, namely a picture of cat with the caption 'cat', before the child encounters a real cat, *world-relating*? If a human child could know this world via and only via second-hand information, could a machine know this world via and only via second-hand information? And if we acknowledge second-hand information as world-relating, we may program to link pictures (technically are symbols as well) to symbols of language and get the condition of 'world-relating' fulfilled, despite hundreds of cat pictures may have to be kept. Note that in Searle's Chinese Room, there are purely formal programs and symbols and nothing else. And in our revised room, there are purely formal programs and symbols and nothing else as well. Only that in our room, there is a second-hand information package serving as world-relating information. Would our room make strong AI stand?

Needless to say, such a solution is far from satisfactory. Not only the Chinese Room but also what has been revealed by the Turing test demand a human level function of understanding in the simulation of natural language processing.

ELIZA, an actual computer chat program past the Turing test, is considered as not intelligent (Block, 1995a). The key strategies employed by ELIZA, as identified by Block, include 'looks for "keywords" on a list supplied by the programmer' and 'transform "you" into "I" '. The comment below on such programs is critical:

> Every clever remark [such a machine] produces was specifically thought of by the programmers as a response to the previous remark of the judge in the context of the previous conversation. (*ibid.*)

The problem is indeed fundamental, as we human beings do not make changes to finite sentences based on finite rules. Instead, we use finite rules and potentially infinite vocabulary to generate infinite sentences.

Borrowing Block's words, to link 'past-oriented intentionality' and 'future-oriented intelligence' (*ibid.*), to link known information 'about' the actual world and doing including saying, we want a capability of, for instance, recognizing a piece of information as new, mentally questioning it, registering it as known, and processing it as a known fact thereafter. So a function of understanding would have to reversely infer the category to which an unknown element belongs. In other words, it is a chain in which one receives auditory/visual information; one understands that the information is about X as an unknown element; one understands what is an X as one has a known category to include X and one believes X should have some properties shared by those belong to the category; one picks one possible

property of X and continues one's conversation.

For instance, suppose this is the first time you listen to your neighbor talking about his cat Ulysses. So 'Ulysses' the name becomes X to you. You know the name is about a cat, and a cat should be an animal that has such-and-such features. So you may pick one possible property (an animal ages) and ask 'how old is Ulysses?'. Or, you know that 'Ulysses' is the title of the novel written by the Irish writer, so you may turn your conversation to 'oh, so you are a fan of Joyce?'.

Reaching second-hand information package and registering new information could only be some parts of our understanding function. Solving the problem of understanding would further require revealing and simulating capabilities such as responding to metaphor. I suspect that metaphor may demand an analysis upon analogical structure, which could be visual than verbal. We might eventually find we need to upgrade understanding from a single function to a module.

4.2 Thoughts, Emotions, Awareness

> Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain-that is, not only write it but know that it had written it. …

Above is a part of Geoffrey Jefferson's oration in 1949 as an objection to the thinking machine. Turing quoted it in his 1950 paper and considered it as 'the argument from consciousness'. Despite indicating a simulation of the working language (or mechanism, or whatever other terms) of the human brain, Turing seems to believe that satisfactory performance in language outputting ensures a kind of thinking. By quoting Jefferson, Turing was rather refuting the 'solipsist point of view'.

Imagine such a sonnet writing program named Rik built with a minimal chain involving the actual world. Set a rose as our object to be perceived. Rik 1) perceives the rose; 2) identifies the rose as a rose; 3) generates an appreciation of beauty; 4) generates an impulsion of writing; 5) thinks about the rose, gets some idea of writing; 6) organizes words based not only on grammar and vocabulary but also on the specifically learned rules of sonnet; 7) outputs the sonnet; 8) has a function *Q* monitoring a part of the steps from the beginning to the end.

Interestingly, although simulating functions of thoughts and of emotions could be highly challenging, with the hypothesized function *Q*, thoughts and emotions *felt* (or, experienced), and *know* that something has been done, are not as challenging. Even when there is no proof for function *Q*, AI could still create such a function that just reviews the outputs from the function of thinking and from the function of emotion.

In the minimal chain, step 5) needs only a bit from the actual world as the background information. Fill the program with only a few propositions about a rose, and link the propositions to perceived information. This makes the aboutness more decisive than that in our revised Chinese Room. The program need some capabilities of reasoning at this step so that it can reorganize the known propositions. For step 6), still consider dictionary Mary, who has been unfortunately imprisoned in a room since she has been born, and almost all the words and visual information she learnt are indirectly linked to the actual world. This is very much the case of current sonnet writing programs, only that current programs typically has

not a single word related to the actual world or to second-hand visual representations. Rik avoids such a situation with the stated solution in step 5). So we are going to have a program lacking directly experienced information and imagination but erudite. If we further minimize step 6), that is, to fill our program with a very small word bank but still quite a full set of grammar, then we have a program lacking imagination and highly ignorant.

Frankly, I doubt based on such minimized conditions if we really can have a program to output a *sonnet* as the program may be out of words for rhyme. But it is certainly possible for it to under such conditions output a poem like 'Oh, you beautiful rose! Oh, how beautiful you are! Oh, you are beautiful!' So Rik is an ignorant, stupid, low in performance program, but it is related to the actual world, has thoughts and emotions felt, and knows what has been done. Would you, as a human being, consider it capable of thinking, and consider it equal to you?

5. Summary

This article first explored the relation between the color in the physical world and the color we see. The discussion shows that the color of objects is not dependent on human minds, while the color we see is. Despite the color we see is mind-dependent, painting and protective coloring ensure us that views on the raw color of objects in the actual world not only are largely shared among human beings but also achieve cross-species agreements.

The article further proposed a dual-field with function *Q* hypothesis of consciousness (DFFQ). DFFQ claims a single function *Q* that could explain the 'first-person point of view' and so could explain the hard problem of consciousness: consciousness could be like an operating system coordinated by function *Q*. Whether there is such a function *Q* in our mind or not could be scientifically testified. DFFQ also says that we might have two different kinds of functions, namely Basic Functions and Coordination Functions, designed to solve various problems working in two dynamic fields in our brain.

Although DFFQ is yet a hypothesis, it is theoretically possible to apply it in artificial intelligence and keep strong AI true. Some practices in AI have been on attention and on memory (Hassabis,et al., 2017), function Q might help in turning such practices into one chain. On the other hand, we might want to carefully evaluate the moves we may take. The robot Landaree in Isaac Asimov's *Robots and Empire* is seemingly intelligent. However, a closer look would reveal the simple and insane structure of her mind: identify accents, mark 'foreign' accents carriers as not human beings, launch fatal attacks against those marked not human beings. We certainly want robots to be intelligent some day, and we may also want them to be *intelligent* in the other sense. Spending a bit of our intelligence upon such issues of artificial intelligence would not be a waste of time.

Declaration

The MIT online course '*Minds and Machines*' *has inspired all the discussions here in this article. However, in case the view in this unsupervised article offenses the academic community in any way, as an online attendant after the course has been archived, the author takes full responsibility.*

For some reason, the author mainly consulted the references introduced in the course. The article thus faces a very high risk of stating something that has already been stated. In such cases, please directly complain to the author.

References

Below are excerpts introduced in the MIT online course '*Minds and Machines'* at edX*:*
[Online]*https://courses.edx.org/courses/course-v1:MITx+24.09x+3T2015/info* [30 July 2017]
Block, N. (1995a) The mind as the software of the brain, in Osherson, D.N., Gleitman, L., Kosslyn, S.M., Smith, S. & Sternberg, S. (eds.): *An Invitation to Cognitive Science*, Cambridge: MIT Press.
Chalmers, D. (2015) The hard problem of consciousness, in Rosen, G., Byrne, A., Cohen, J. & Shiffrin, S. (eds.): *The Norton Introduction to Philosophy*, Norton.
Hardin, C.L. (1984) Are 'Scientific' Objects Coloured?, *Mind*, 93, pp.491-500.
Jackson, F. (1982) Epiphenomenal qualia, *Philosophical Quarterly*, 32, pp.127-136.
Nagel, T. (1974) What is it like to be a bat?, *Philosophical Review*, 83, pp.435-450.
Place, U.T. (1956) Is consciousness a brain process?, *British Journal of Psychology*, 47, pp. 44-50.
Ryle, G. (1949) *The Concept of Mind*, Hutchinson.
Searle, J.R. (1980) Minds, brains, and programs, *Behavioral and Brain Sciences*, 3, pp.417-24.
Searle, J.R. (1990a) Is the brain's mind a computer program?, *Scientific American*, 262, pp.26-31.
Smart, J. J.C. (1959) Sensations and brain processes, *Philosophical Review*, 68, pp.141-56.
Turing, A. (1950) Computing machinery and intelligence, *Mind*, 59, pp.433-60.
Tye, M. (2015) The puzzle of transparency, in Rosen, G., Byrne, A., Cohen, J. & Shiffrin, S. (eds.): *The Norton Introduction to Philosophy*, Norton.
Course slides 2:6, the MIT online course *Minds and Machines.*

Other Sources

Block, N. (1995b), On a confusion about a function of consciousness, *Behavioral and Brain Sciences*, 18, pp.227-287.
Gazzaniga, M.S. (2011) *Who's in Charge? Free Will and the Science of the Brain*, Ecco.
Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. (2017) Neuroscience-Inspired Artificial Intelligence, *Neuron*, 95(2), pp.245–258.
Mason, P. (2014) *Week 3, Language Circuits,Understanding the Brain: The Neurobiology of Everyday Life*, The University of Chicago, Coursera,[Online]
https://www.coursera.org/learn/neurobiology/lecture/dWdxw/language-circuits [29 July 2017]
Mason, P. (2014) *Week 4, Learning to see, Understanding the Brain: The Neurobiology of Everyday Life*, The University of Chicago, Coursera, [Online]
https://www.coursera.org/learn/neurobiology/lecture/oXPuj/learning-to-see [29 July 2017]
Not a panpsychist but an emergentist?, Conscious Entities,
[Online] http://www.consciousentities.com/2014/01/not-a-panpsychist-but-an-emergentist/ [29 July 2017]
Oh,Phi! , Conscious Entities, [Online] https://www.consciousentities.com/2012/10/oh-phi/ [29 July 2017]
Searle, J.R. (1990b), Consciousness, explanatory inversion and cognitive science, *Behavioral and Brain Sciences*, 13, pp.585–642.
Van Gulick, R. (2016) Consciousness, in Zalta, E.N. (eds.) *The Stanford Encyclopedia of Philosophy*, [Online] https://plato.stanford.edu/archives/win2016/entries/consciousness/ [30 July 2017]