# Chapter 12

# Consciousness

*A difficult but important idea to get used to is that the brain has not access to ground truth about the world and people in it but instead runs a model of these things. Most of this model is driven by previous experience but at least for humans, an additional component is the ability to simulate situations in the future. Naturally to be successful at this has enormous survival value. However to run this simulation requires additional mechanisms that are not covered by purely memory-driven algorithms. These mechanisms are the candidates for a substrate that produces the feeling of conscious experience.*

Before we try to pin down some attributes of consciousness, it might be helpful to visit the feeling of being unconscious. We have all been unconscious in a dreamless sleep. Nonetheless it is a very gentle unconsciousness because, as you know, while you are asleep the brain has a lot to do to save the experiences that were selected form the last period you were awake. In addition, before you went to sleep, you had expectations about what things would be like when you awakened, and for the most part these are met. Perhaps a better experience of unconsciousness is obtained by those of us who have been under anesthetic during a medical operation. In these cases the drift into unconsciousness and awakening is abrupt and the interim experiences are a blank. Without the aftereffects of whatever procedure we have had, the time we were under the anesthetic would be just seamlessly missing.

The experience of being conscious and not conscious, the latter as measured when consciousness returns, is easy to relate to because we all have it. But while we all have the feeling of being conscious, to try and jump into

a satisfactory explanation of its holism is fraught with difficulty. The main reason is that 'consciousness' is another one of our summarizing words that tags a lot of underlying structure. Thus the job of this chapter is to deconstruct the experience of consciousness and examine its components and their respective functions together with the rationales for those functions. You have to be prepared for a disappointing experience. Think of a car. One can take it completely apart and layout all its pieces on the garage floor with their attendant functions labeled, but the dissected and labeled 'car' is not the same thing as the assembled vehicle.

## 12.1 Having a Model

In beginning the deconstruction of consciousness, the most important of its component concepts is that of a model. The brain's programs of course do not have access to all the information about the world, but only receive a small shadow of that information through the process of sensing and acting in it. Thus the programs have to estimate the essential state information about the world in order to direct these programs. The consequences of working with state estimates rather that the real thing are profound and we can only really appreciate them when something goes wrong as when a brain is injured.

A spectacular divergence of a brain's estimate and ground truth occurs with phantom limb patients, explored by the psychologist and physician V.S. Ramachandran[5]. For most people who have the misfortune to lose a limb in an accident, that's it. The limb is gone and they learn to adjust to life without it. But for a few people have an unusual condition where they swear the limb is still there. They of course have very conflicting data in that they can look at the empty space where the limb used to be. Yet they still are quite convinced that they have it. You would think the issue would be settled when Ramachandran asks them to do something with the limb, like pick up a cup, yet the patients are nonplussed and come up with an excuse like "Oh, I can't do that right now because my arm is too tired."

Ramachandran's hypothesis is that what has happened is that the cortical circuitry handles the face is near that which used to handle the arm and gets co-opted into service by the programs that were running the arm. Thus the neural circuit that handled the arm is satisfied. It needed inputs and it has them. Naturally there has to be some evidence combination that

reconciles the visual input with the ersatz haptic input, but we can assume this happens and that the latter trumps the former. If you think in terms of your phenomenological experience of vision this is hard to swallow, but if you think in terms of tests that have varying results it becomes much easier.

The astonishing phantom limb data contain two very important concepts. One is that the model running the arm that uses the cortex is in a certain state and at a lower level of abstraction than that of conscious report. Furthermore conscious report has no alternative but to summarize the situation down in the engine room as best it can. Hence the rationalization and equivocation. The other point is that conscious report, interrogating this less abstract state uses time and necessarily comes after that state is generated. The idea that consciousness comes after subconscious processing we have seen in the discussion of emotions, but also has appeared in numerous other contexts [3, 4].

The idea of running a model is not limited to just injured brains but of course is a general feature of working brains as well. A nice demonstration comes from experiments done by [6]. Subjects viewing a display saw two circles move behind a rectangular occluding surface as shown in Figure 12.1. There were two different conditions. In one as the circles met behind the surface - hidden from the observers - there was the sound of a 'click.' When this happened, subjects perceived the circle on the right as having collided with the circle on the left, bouncing off it and emerging on the same side. The left circle behaved oppositely. However without the click the perception was reversed. The circles were seem to pass by one another and emerge on opposite sides. This is a stunning demonstration as is reveals clearly that the brain has embedded the visual and auditory in some kind of model. The visual histories are identical in both cases; its just that in the first the sound is associated with the physical act of colliding and changes the course of the internal simulation.

The ability of using models to interpret the world is so seamless and effective that we do not readily appreciate that it is acquired during child development in a lengthy staged process that spans the years from birth to puberty. The study of this process has been enormously productive and has revealed many features of the different stages. We will just limit ourselves to a single observation that conveys some of the complexities involved.

Experimenters showed two and a half year olds a drawing of a living room scene with a teddy bear in the drawing behind a couch. When these children are subsequently taken into the real room for which the drawing was a model,

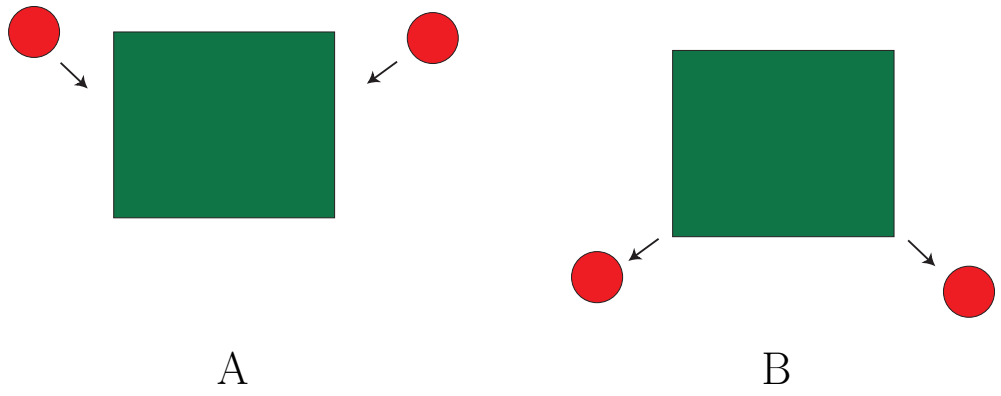



Figure 12.1: Two frames from an experiment to test perception A) A moment before two moving circles disappear behind an occluding surface B) A moment after they reappear on the other side.

they have no trouble finding the teddy bear. They have understood that the drawing represents the room and make the correspondences necessary to locate the teddy bear. Its what they cannot do that is especially interesting. Children of the same age are shown a doll house that contains a replica of the room that they will go to to find the teddy bear. In the replica the bear is placed behind the couch. But when these children are taken into the real room they are clueless. Somehow they cannot appreciate that something already in the real three-dimensional world(but is just smaller) can also be a model for something else in the three dimensional world. But at the same time they can understand that something that is not three-dimensional at all and just contains a facsimiles of things in the three-dimensional world is a useful representation of that world.

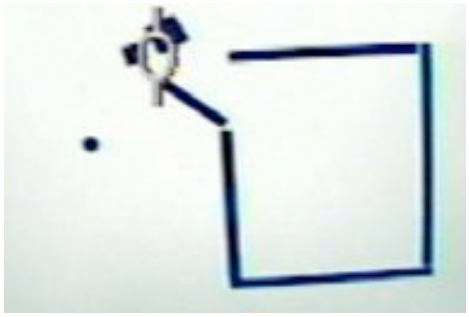
In some sense this result might seem counterintuitive to us. Why does a child not find the three-dimensional problem easier to solve? As adult observers we can appreciate that the difference is just one of scale, and once that simple notion is understood, the problem is trivial. However the genes have a different take on the problem, since the abstract encoding where the picture codes the problem is easier to solve. The paper features tip off the child that the information on the paper contains symbols for something real.

## 12.2 Agency

Up to this point there have been two crucial ideas, the first being that the brain runs models and the second being those models prefer abstract currencies in their depictions. As you saw, the ability to abstract from two-dimensional images is highly developed in adults and we use it effortlessly.

What is also developed is the use of abstraction for characterizing agency. On an evolutionary time scale, anything that moves substantially is most likely to be an animal. Thus this programming extends effortlessly to symbolic representations that move. As an example, we showed subjects movies of simple diagrams with movies of simple moving tokens represented with filled squares, and circles. Figure 12.2 shows frames from one such movie. The movie contains only geometric shapes in motion, but when asked to describe such scenes, subjects vividly bring the scenes to life attributing sex and elaborate motives to the moving tokens. In the figure the eye fixation point is also indicated as a white marker. The box is naturally a house seen from above. The small figures are seen as a pair of animates that have to deal with the large square. What is interesting is that normal subjects inevitably have huge overlaps in their descriptions of the scene. They see the movie as one of conflict between the large square and the other two figures. In the frames shown in Figure 12.2 note that the depiction is consistently from the viewpoint of the small figures suggesting an identification with their point of view.

After reviewing the evidence for how we do test the environment in Chapter 6, perhaps we should not be surprised that we just use the barest of elements to define complex events. We are not worried by the absence of evidence of enfleshed characters. The story is a little different when there is positive evidence for details that are not quite right though. Jack Loomis makes studies of human interaction with virtual human figures, or avatars, in virtual reality(VR) environments. He is after just this question. What cues about the virtual figures need to be present for the subjects to treat them as real? Subjects wear a head mounted display and see the virtual figures that may be engaged in various activities or just standing around. In one experiment he instructed subjects to remember a number printed on the front of the avatar and a name printed on the back[1]. Since the avatar was standing still, to do this they had to walk around it. Loomis had two different ways of describing the avatar; in some runs observers in VR were led to think that the avatars w ere driven by humans and in other runs they



"The Big Square (BS) confronts the Little Square (LS)"

"The BS chases the Little Circle (LC) into the house and confronts him"

"The LC runs out of the house and the LS shuts the door behind them"

"The BS is moving very fast; he's chasing them"

Figure 12.2: Four frames from a movie that has a simple senario involving a circle and two squares. Normal subjects effortlessly attribute agency to the moving geometric figures The eye position is shown on each frame.

were led to think that they were driven by a computer. The interesting result was that the subjects were very sensitive to these details behaved towards the avatars differently under the different assumptions. When non-human driven subjects did not respect the avatars personal space in the course of getting the name/number information. But when the avatars were thought to be driven with great fidelity, the subjects gave them agency and respected their personal space, walking around them at a close but polite distance. Figure 12.3B shows another variant. Subjects gave staring avatars a wider berth than non starers. Even though that state of the avatars was known to be identical in every other respect.

A B

Figure 12.3: A) A walking environment for Virtiual Reality experiments. B) Part of the experiment designed by [1] to test agency.

In another experiment subjects are playing blackjack in virtual reality for a while when two avatar players appear beside them and also start playing. Since the subject has been playing for a while he has established a risk level in terms of the average amount that is bet on each hand. It turns out that social pressures are such that in a group you like to bet a little more than the average of the group, because being male risk taker has status. The delightful result is that when the avatars appear and start betting more than you, you'll raise your bet to impress them. Remember they are not real, just graphics figures!

## 12.3 Autism and mindblindness

It rather astonishing that we endow avatars with an agency. We know that they are not humans of course, yet in behaviors we accord them human respect. However now that you have the picture of how the brain handles things in the cortex, it is perhaps a bit easier to appreciate; The brain uses tests to index behaviors, and the avatars pass enough of them to be treated with human respect.

The extensive social structures that we have to navigate, together with their accompanying linguistic demands would seem to make it impossible that we could ever lose this sense of agency, yet in the special case of autism, it appears that that is exactly what happens. Severe autistics have enormous problems reading the minds of other people. Baden-Cohen whose life interest is in studying autistics, has termed this "mindblindness."[2]

The classical experiment, done with young children has an autistic in a room standing in front of two boxes, one of which contains a candy box of "Smarties." The child's friend Fred leaves the room and the autistic sees the experimenter take the candy from its box, lets say Box A and put it in the other box, Box B. The autistic is then asked: "When Fred comes back, which box will he say has the candy?" Of course he should say Box A because Fred did not see the transfer and would infer that the candy is in its original box. But the interesting observation is that the response is Box B. Why? The thought is that the autistic cannot make a distinction between what he knows and what Fred knows. Since he saw the transfer and knows that the candy is Box B, then Fred must think so too.

From a computational perspective, having trouble with agency in this way is not a big leap. We have seen that the important features of our own agency our intimately tied up with our body machinery. When we experience emotions, we are interpreting our momentary body state. In this light, appreciating the agency of others means that we have to simulate them using our hardware. Thus we have to allow our readouts of their state to direct our internal simulation. Even normals doing this often get it wrong as we can all testify. At any rate you can appreciate that this might be a technically difficult fragile program that could be damaged. Amazingly, autism is due to a genetic abnormality. The genes seem remote in the sense that they work at a very low level of abstraction to make proteins that construct the phenotype. Yet here the long arm of the gene reaches up to produce a profound and seemly subtle defect.

Baden-Cohen's thesis is that a necessary precursor in the normal developmental pathway is that a child learn to use a caregiver's eye movements as a pointing tool. The caregiver has a pre-linguistic way of instructing the child just by looking at things. If a child fails to pick up on this indicator it signals the beginning of trouble distinguishing the intentions of another person and his or her own. There is lots of evidence to support this idea including some from our own lab. Chen Yu has shown that human subjects can learn words from a foreign language much more easily when they see eye movements. In his experiments, humans listened to a child's story read in Mandarin. None of the subjects had any experience with Mandarin. There were two conditions: in the first they heard the story read and saw the pages of the story book turned at the appropriate points, in the second additionally they saw an indicator of the reader's eye gaze position on the text from moment to moment as the story was read. Afterwards the subjects heard a animal's Mandarin code and had to pick one of five possible animals that it referred to. Figure 12.4B shows the result. Subjects who had seen the gaze position during the reading of the story did spectacularly better than those who had not. Subjects who had seen gaze got over 60 % of the tests correct while subjects who had not, performed no better than guessing.

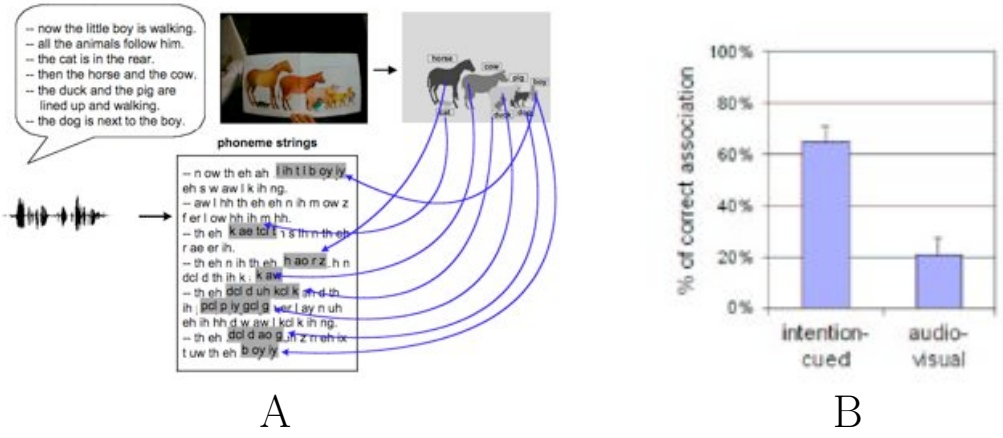



Figure 12.4: A)A basic problem in learning a language is that the learner has to associate parts of the speech stream with referents B) Experiments show that human adult language learners are much better at this when they can see the teacher's eye gaze position.

Needless to say having autism leads to all kinds of challenges for the

person who has this difficulty, but the point here is about agency. A concept that we take for granted does not come without a struggle and in the case of autism can be sidetracked. You should also take notice that this deficit is not necessarily about intelligence as autistics can have very strong intellects. Its about agency. Interestingly, at least some chimps do not have trouble with the Smarties task. In an experiment with chimps done at Harvard, experimenters arranged a viewing situation with two humans. When food was put in one of two places, one of two humans was able to see where it went and another was not. In the next phase both humans indicated where the food was, pointing to different locations and the chimp subject had to choose the correct location. The crucial question was whether the chimps could use which of the two humans had the correct knowledge of the location from having watched the setup senario. Although some chimps were clueless, others caught on and were able to get the right location consistently.

## 12.4 Conscious Will

If there is a jewel in the crown of consciousness, its our so-called free will. We have the distinct sensation that we can choose among actions. I have the sensation now that I could go into the kitchen and make either coffee or tea. It seems that I make up my mind as to which one I want. But the thesis of this book is that its all computation. So in that light we ask what does it mean in a program to make choices?

Making choices in the exercise of our free will is not always so easy. We just saw that phantom limb patients do not have the ability to believe that their arm is amputated, despite huge amounts of evidence. In virtual reality laboratories you can have people wear a head mounted display and shown them a huge virtual pit in front of them. If you ask them to step into it, a large percentage of subjects refuse. They know that the ground in front of them is the solid laboratory floor, yet the visual evidence that it is not is compelling. So at best, free will consists of weighing evidence in the choice making. If there is too much evidence mediating one choice, others are unavailable. I might muse that I have the free will to rob a convenience store, and perhaps we could create circumstances such that it would be an option, but the fact is that under any normal set of circumstances I would not be able to do it. I do not have the freedom or free will to make such a choice.

Nonetheless, despite these arguments at least some readers are still think-

ing to themselves: "maybe that is the way your brain works, but I still feel that I have free will." Daniel Wegner has tackled precisely this question in a wonderful experiment that is able to separate when we feel we do something from when we actually do it, not an easy thing to do. The experiment makes use of a setup that is reminiscent of an old parlor game played with a special board - the Ouija board, so we will take a little detour to introduce.

The Ouija board is a parlor game from a more sedate time. The board itself is a hardwood surface with "yes" and "no" alternatives marked on alternate sides of the board. Two people sit with the board on their laps and manipulate a *planchette*. This object is a raised surface that they can comfortably grasp from either side. The planchette is on pedestals designed to make it slide easily on the board. The players cooperatively slide the planchette about on the board. One player thinks of a question and then the subsequent motion of the planchette is supposed to guide it towards the answer.

What makes the Ouija board fun is that for some pairs, there is a compelling experience that the planchette is moving randomly and is not under the control - or wills - of the two players. Thus the answers to the questions seem to come from an external source.

Incidentally from what we learned in the previous chapter, we can make a suggestion as to why this works - when it does. The players are trying to be cooperative since its not fun if the planchette moves deterministically. What is means is that they are trying to be random. Since each persons the movements are more naturally deterministic, it could be the case that generating random movements is easier in a the cooperative game. The players sense the movement of the other player and try to generate a movement from their since that is not in that direction, a fairly easy thing to do. At the conscious level the movement of the board seems random because, owing to the success of coupled two-player game, it is random.

In the Wegner experiment, two people manipulate a small planchette-like structure over a table. The planchette controls a computer cursor that glides over a cluttered scene containing many common objects. The participants hear audio signals that instruct when they are to stop moving the cursor and place it over an object. The experimental cleverness is that one of the participants is actually a confederate who hears precise instructions as to when to stop moving the cursor. A trial consists of 30 seconds of movement followed by 10 seconds of music indicating that they should make a stop. The subject heard words over the headphones that were billed as there to

provide a distraction. The confederate was of course supposed to be hearing distracting words also, but instead heard instructions on what to do. The interesting subject of the trials were forced stops. On these the confederate was instructed to move towards a specific object and given a countdown with which to use to stop on that object. The purpose for all this structure was that it allowed Wegner to play a priming word to the subject at specific times related to when the confederate stopped. Thus the subject thought she was stopping but in fact that was being controlled by the confederate. Both people were asked to rate the all the stops as to intentional they were on a 14 point scale ranging from "I allowed the stop to happen to "I intended to make the stop." The priming word in the forced stop was at 30, 5 and one second before the action and 1 second after the action.

The results are shown in Figure 12.6B. The interval rated the highest as to intentionality was one second before the action was made. Keep in mind that for none of these data is the subject actually controlling the cursor. One second after gets a low score as does 30 seconds prior. Wegner's model for this can be seen in Figure 12.6. The idea again appeals to the abstraction gulf between the machinery that is directing the selection of actions and that which is privy to conscious report. In this case, hearing the word primes the low level circuitry that actually generates the action. In reinforcement learning parlance, priming is part of the indexing state that will guide the action and stopping is action selection. So like the phantom limb patients, here the normal experience of conscious will can be interpreted as the product of running an internal model. In this ingenious experiment Wegner is able to manipulate the particulars of that model in his subject. The idea is that conscious will depends on a timing relationship between the computation of the state that guides the action and the execution of the action itself. The experiments suggest that normally this is about one second or less before the action occurs.

## 12.5 Simulation

At this point you have seen the basic components, so that its time to put them together to obtain a description of consciousness. To do this lets revisit unconscious behavior.

To compensate for slow (compared to silicon) circuitry, the brain stores behaviors in tabular form. The cortex is a vast storehouse of millions of

A

B

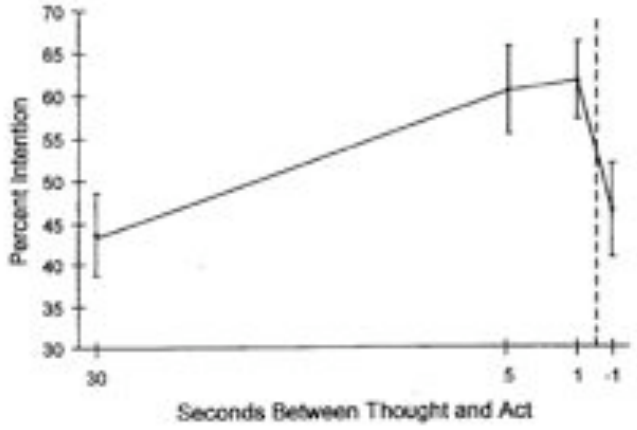



Figure 12.5: A) The confederate and subject in the I Spy experiment moving an ouija-like cursor to targets on a screen B) The ratings of willfulness produced by the subject pool depend on time between when a prime word was heard and the actual stop of the cursor. [7] [Permission Pending]

prescriptions for the tiniest aspect of behavior that includes seeing, speaking and feeling. These behaviors are rote; we don't think about them when we execute them. Think of driving a car along a familiar route. You probably are conscious of times when you have been unconscious, that is you drove for miles without being aware in any deep sense of transiting the route. When you finally become aware you can wonder whether you ran a red light or stop sign but you are clueless. Of course you didn't. Your detailed set of programs for driving that familiar stretch directed eye movements, head and and foot movements, stopped at red lights, started at the onset of green lights, signaled turns and almost countless other small but necessary tasks. This example is just one of the myriad of behaviors that your brain stores.

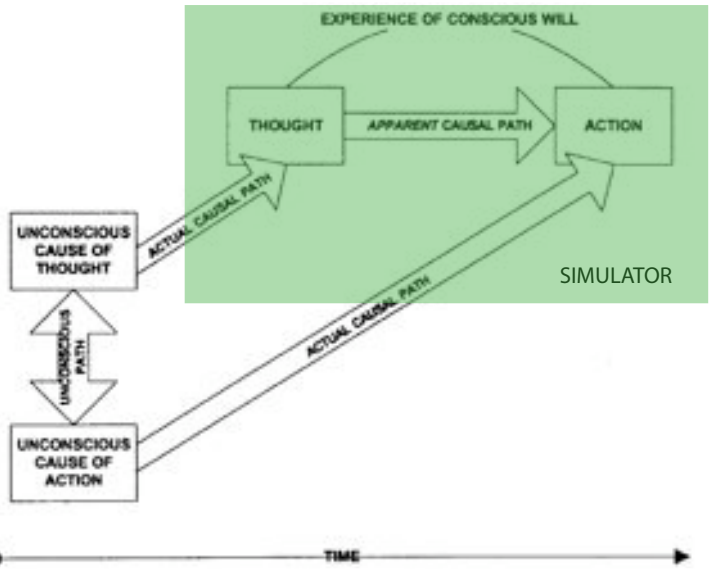



Figure 12.6: In Wegner's model the illusion of conscious will is produced as a byproduct of an expected half-second delay between the appearance of the representation of the action and the execution of the action. [Permission Pending]

In fact that's what its for; to execute a stored repository of behaviors that allow it to accomplish the goals laid in via millions of years of evolutionary programming. Furthermore those programs are constantly being fine-tuned by loads of repetitions. But none of this is conscious.

So what is the role of consciousness? Almost everyone agrees that the number one value of having a model is the ability to predict the future. Since the brain runs models, its easy to imagine a reward system that places a great premium on brains that are slightly better at doing this. However predicting the future is not easy because it has not happened yet. Thus there are two crucial aspects that have to be dealt with.

1. The model must be able to run ahead of the present, constructing representations of possible worlds. This ability introduces a difficulty in that modeling the present and near future is easier because ether is a vast source of sensory and motor cues of just how the present *is*. When the model is projected in the future, it looses access to this vast sanity check.

2. The model must deal with potentially one-off situations that are not explicitly represented in memory. There are some situations that one could expect the memory to handle and those are where the exact situation has not been seen but it can be interpolated from behaviors that have been seen. However after we remove these from consideration there will always be situations that have not been encountered. They might have a component that can be looked up in the table, but there are still some unique elements left over.

These two elements in conjunction mean that the programs that can run into the future have to have special features to deal with them. Lets run the risk of producing another jargon word that needs to be deconstructed and call these programs collectively *the simulator*. The simulator's primary value is that can run in the future, but it can also be tripped by either of the two conditions just mentioned. For example you can be doing something in the present and stumble into a novel aspect not covered by the tables. The simulator kicks in and can run the tables with hypothetical assignments.

Lets imagine your ancestor of 10,000 years ago hunting the world's mammals. He forms a plan: he will distract the mammal while the rest of the group pelts it with spears and or stones. In this construction, it is vital to not get confused between the distractor and the attackers. The brain has to represent both parties with their hypothetical duties. In the brain it is very unlikely that it is done with a few 'grandmother' neurons in a specific place. As you saw previously, a much more likely candidate is the emotional system. For each person, the cortical imprint of the body's different functions provides a signature that means 'ME.' From this perspective, you can see the technical demands of tagging a person that is not yourself. You have the mechanism: you could just attribute the signature to another, but that would leave you severely autistic. What you have to do is mark the differences in that signature in a way that you have created another distinct individual. In effect the simulator is doing a second-order simulation. You are not just simulating yourself in the future where you can tag items with your signature, but you have to abstract the signature itself to create an independent person 'symbol.'

One very important concept that can help here is that of working memory. Working memory is needed precisely in the case where an unfamiliar referent has to be spliced into a familiar program. Thus it is a correlate of the running of the simulator. It helps keep track of the fact that you have left the realm

of every day existence and are running in hypothetical mode. The simulator uses this structure to adjust the value of itself - the plan ahead program - against the standard background cacophony of routinized behaviors.

# Bibliography


[1] Jeremy N. Bailenson, Jim Blasovich, Andrew C. Beal, and Jack M. Loomis. Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin*, 2003.

[2] Simon Baron-Cohen. *Mindblindness An Essay on Autism and Theory of Mind*. MIT Press, 1997.

[3] Bejamin Libet. *Mind Time the temporal factor in consciousness*. Harvard University Press, 2004.

[4] Tor Norretranders. *The User Illusion*. Pengiun Books, 1999.

[5] V. S. Ramachandran and Sandra Blakesee. *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. Quill William Morrow, 1998.

[6] Robert Sekuler, Allison B. Sekuler, and Renee Lau. Sound alters visual motion perception. *Nature*, 1997.

[7] D. M. Wegner. *The Illusion of Conscious Will*. MIT Press, 2002.