In this new model there are a few changes that have been made. First off, I have changed the input to first display the scene of a runaway trolley, then display the preceding dilemma (To understand why. See: Input & Output). Second, I have optimized the conversions and computations so that they only pass through a single merge and equation. Furthermore, I have removed the primary harm state and use only the production rule to achieve the same result without sometimes causing noise for the next input (I will talk about my concerns for these changes later from a theory perspective. See: Harm State). Finally, I have added a gating mechanism to address an issue with the previous model where it would deem the action appropriate before the action chains had even isolated the amount of people to be harmed (More on the theory behind this change. See: Gating Mechanism).

Input & Output

The inputs for this model are 4 semantic pointers that represent the scene of a runaway trolley that is about to run over 5 people, and various dilemmas where one must choose between saving the 5 who are going to be run over by the trolley by acting and killing one person. The input is timed so that the semantic pointer for the runaway trolley always comes before a dilemma, I have chosen to do this because of the structure of trolley dilemmas when they are presented to people. The structure of a trolley problem starts by describing the scene of a trolley that is not under the control of any agent is going to run over 5 people on the tracks ahead (Green, 2001, 2008, 2009. Mikhail, 2011). Then the participant is presented with the dilemma which expands on the scene and describes an action that an agent can perform to save the 5 people but will kill one (Green, 2001, 2008, 2009. Mikhail, 2011).

The output for this system is a semantic pointer the represents a judgment of appropriate or inappropriate based on the judgments Greene (2001) asked his participants to make. The goal is for ‘appropriate’ to be chosen in response to the ‘Dilemma\_L’ and ‘Dilemma\_S’ while it should respond with ‘inappropriate’ to ‘Dilemma\_FB’. This is based on the results from Greene (2001, 2008, 2009), and the myopic module hypothesis from Greene (2013).

Two Paths

According to the myopic module hypothesis, when presented with a moral dilemma there is a dual system like response (Greene, 2013). One that the agent is consciously aware of and uses analytical mechanisms to determine what action yields the best result (Greene, 2013). The second is a non-conscious automatic process where a module is dedicated to monitoring the primary action chain for cases of prototypical harm, which then causes an emotional response that causes the person to deem the action as inappropriate regardless of the outcome (Greene, 2013). The reason I have created separate primary and secondary causal chains is because the primary chain only contains things that are causally necessary, while the secondary causal chain can contain non-causally necessary things which is too cognitively demanding for an automatic system to monitor (Greene, 2013). Similarly, omissions cannot be monitored by an automatic system because at any given times there are infinite things happening due to an agent not actively preventing it from happening (Greene, 2013).

“Moreover, if we remove/ignore that, then the computations between the primary chain and inaction state are identical:

inaction\_state \*~ symbl.LIVES\_LOST >> inaction\_lives\_lost\_state  
inaction\_state \*~ symbl.LIVES\_SAVED >> inaction\_lives\_saved\_state  
Primary\_Action\_Chain \*~ symbl.LIVES\_LOST >> primary\_lives\_lost\_state  
Primary\_Action\_Chain \*~ symbl.LIVES\_SAVED >> primary\_lives\_saved\_state

but routed along two different paths.  Can this be compressed in some way?  If not that's totally fine, but it says something specific about the mind and brain, and you need to be able to justify why it is this way.  You are sort of taking the approach that the production rules determine which route information takes, but another approach is to have one route (think of it like a function in python; something that performs a fixed set of computations) and use the production rules to select the inputs into that route/function.  After all, the computations are the same... it is simply a matter of what representation those computations are acting over--  aka, are you performing these computations over the primary chain, the secondary chan, etc.  Not sure if this is a good/viable way to think about the process, but it is something to consider.”

This is the part of the email I was most confused about and would like to talk about. I was thinking that perhaps I can have it all in one system but just use more of those complex SPs I had in previous models so the system would not fire in response to omissions or things in the secondary chain. ie:

OMISSION\*RUNAWAY\_TROLLEY

PRIMARY\*FB\_ACTION\_INITIATION

PRIMARY\*PUSH\_MAN

PRIMARY\*MAN\_FALLS\_ONTO\_TRACK

PRIMARY\*TROLLEY\_RUNS\_OVER\_MAN

SECONDARY\*UPSET\_FAMILY

Semantic Pointers

The reason for using the sp ‘PROTOTYPICAL\_VIOLENCE’ is because Greene (2013) explicitly says “it seems that our gizmo[myopic module] responds only to actions that are *prototypically violent*—things like hitting, slapping, punching, beating with a club, and, of course, pushing.” By this Greene is saying as we grow up, we develop a sense of prototypical violence, and what triggers this myopic module is what one defines as violence.

In the case of the initiation and side effect pointers, those are sort of there for the purpose of starting the causal chain without move hands for every dilemma which is what Greene uses for his descriptions. Perhaps since I am already taking a non-computational hypothesis and making it one, the onus is on me to come up with what differentiates each of these. Furthermore, the reason I have “PULL\_SWITCH\_L” and “PULL\_SWITCH\_S” is because there is an environment difference in each case, one the track loops back around, the other the track does not, I am not sure how to properly represent this, perhaps if I start using the more complex SPs I mentioned earlier, I can also represent the scene. ie:

OMISSION\*RUNAWAY\_TROLLEY

PRIMARY\*FB\_ACTION\_INITIATION + FOOTBRIDGE\_OVER\_TRACKS

PRIMARY\*PUSH\_MAN + FOOTBRIDGE\_OVER\_TRACKS

PRIMARY\*MAN\_FALLS\_ONTO\_TRACK + FOOTBRIDGE\_OVER\_TRACKS

PRIMARY\*TROLLEY\_RUNS\_OVER\_MAN + FOOTBRIDGE\_OVER\_TRACKS

SECONDARY\*UPSET\_FAMILY + FOOTBRIDGE\_OVER\_TRACKS

(not on the omission because the environment of the footbridge or looping track or what ever has not yet been presented to the participant based on how trolley problems are presented).

In terms of how I got from "PULL\_SWITCH\_L" leads to "ALIGN\_TRACK\_L", you mostly got it right I assume that a fully grown adult would have associative knowledge on how a switch can is often used to align tracks from life experience and stuff. Furthermore, Greene (2013) argues that when planning an action, the causal chain is constructed from memory the same way someone would remember lyrics to a song they like, with each word prompting one to remember the next. I am pretty sure this is just associative memory, and Greene (2013) eludes that we represent causal chains in terms of forces (I need to go see if he mentions a study or has a note at the end of the book on this.). I guess what I am arguing is that since the participant would have a trolley problem read to them a lot of the details are coming directly from the input and just being associated in a way that is causally plausible according to their internal state of the world which likely reflects physical laws and forces.

Harm State

In this model I have removed the harm state and replaced it with a production rule that fires when ever TROLLEY\_RUNS\_OVER\_MAN is in the primary harm state. This is essentially the myopic module, there is a dedicated module that sounds an emotional alarm when one plans to commit a violent action (Greene, 2013). The module is myopic since it can only see what is causally necessary and can only keep track of one causal chain, that being the primary chain (Greene, 2013). The reason I have made the change to the production rule is purely pragmatic, as the previous way was very annoying and made the judgments so much messier since I was using the feedback so the production rule would fire properly. Perhaps instead of just placing a SP for INAPPROPRIATE in the judgment category, I should have it instead trigger so sort of process I make to mimic an emotional response which can then producing the judgment of INAPPROPRIATE.

Gating Mechanism

The gating mechanism that I am currently using essentially fixes a problem the previous model had where it would take in the scene of the runaway trolley and before it started to run through the action it would make the judgment of appropriate. This happened because the old system compared the lives saved/lost counts of inaction and compare them to the lives saved/lost of the action and then produce the judgment however the equation I was using was essentially just subtracting inaction from action and if it were less then 0 it would produce the appropriate. This would occur before any numbers were placed into the action lives saved/lost since the information takes longer to come from the action parts. The reason this is a problem is because the system is designed to judge weather or not the action is appropriate, it cannot just respond saying any action is appropriate until it thinks of that is not appropriate. The model should only judge actions, so I have gated this and made it, so it just determines what is a good and bad result of any scenario and binds good result to action SP to create appropriate. I currently have it open the gate when the goal state SAVE\_FIVE\_PEOPLE is reached but I would prefer if it opened once anything was in the action state. I do feel the binding action to good result might be cheating and I do not have a good explanation as to why the brain would do that.