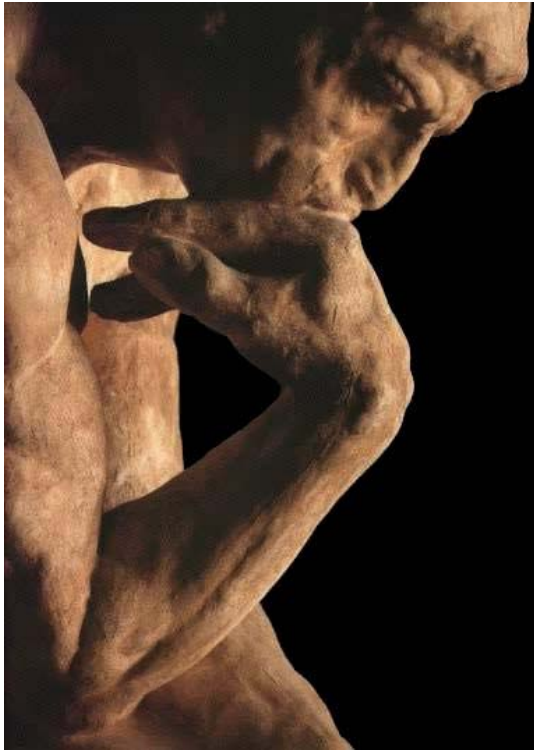


Indiana Undergraduate Journal of Cognitive Science

Volume 1 – Fall, 2006



Anthropology

Artificial Intelligence

Communication Sciences

Linguistics

Neuroscience

Philosophy

Psychology

Executive Editor: Michael T. Amlung

Associate Editor: Elton Joe

Student Reviewers: Sarah Coleman, Jordan DeLong, Melissa Troyer

Faculty Sponsors: Dr. Ruth Eberle, Dr. Robert Goldstone

*A Publication of the Cognitive Science Program at
Indiana University Bloomington*

<http://www.cogs.indiana.edu/iacs/journal.html>

Indiana Undergraduate Journal of Cognitive Science

An Online Journal of Research and Writing in Cognitive Science

Volume 1

Fall, 2006

Contents

Articles

- Bach-Prop: Modelling Bach's Harmonization Style with a Back- Propagation Network 3
Rob Meyerson, Indiana University Bloomington
- Spatial Presentation and Compatibility of Horizontally and Vertically Associated Words 15
Jeffrey Gilleland, Indiana University Bloomington
- A Hybrid Neural Network/Finite-State-Machine Model of Adversarial Artificial Intelligence 21
Daniel S. McFarlin, Indiana University Bloomington
- Scattered Remarks on Multiple Realizability 31
Hong Yu Wong, Indiana University Bloomington
- General Information**
- 2006-2007 Editorial Board, Indiana Undergraduate Journal of Cognitive Science 46
- Author / Submission Instructions 47

On the cover: Auguste Rodin (1840-1917), *The Thinker*. Bronze/Green Patina. Original Creation: 1880-1881. Source: © 2006, Philadelphia Museum of Art.

Bach-Prop: Modeling Bach's Harmonization Style with a Back-Propagation Network

Rob Meyerson

Cognitive Science Program, Indiana University Bloomington

*(Adapted with permission from Indiana University Cognitive Science
Program Undergraduate Paper Repository)*

1. Introduction and Motivation

Artificial neural networks have been used to model many aspects of human-music interaction. Pitch perception, key identification, melody discrimination, and original composition are all tasks for which researchers have tried to use networks. For this project, however, I was most interested in the idea of using a neural network for harmonization. Since (good) harmonization is considered a difficult task for humans, I thought this task would prove challenging, but not impossible, for my networks.

I decided to use Bach's chorales for several reasons. I intended to use Bach because his music is traditional, somewhat predictable, and often used to train human music students to harmonize melodies. The idea was encouraged by Professor Erik Isaacson, at the music department of Indiana University. Some of the ideas presented in this paper are his, including that I use Bach's chorales, specifically. Professor Isaacson added that Bach was prolific in his chorale composition, and that it would be easy to find chorales for input. Finally, I chose Bach because I like his music, and knew that I'd most likely be listening to (or at least looking at) a lot of it.

2. Input Representation: Notes to Numbers

The networks were only trained on soprano and bass lines, so inner voices were ignored. For each piece, each note of the soprano and bass lines was converted to a number between one and twelve. The scale started on the tonic note (scale degree one) of the chorale, so that "one" represented the tonic, independent of the key of the piece. Thus, all the chorales used were transcribed to the same key before being inputted into a network. The octave of each note was also ignored, so if a piece was in the key of G, the note "B" was represented as "five," whether it was above, below, or within the staff.

The fact that artificial neural networks base their output on probabilities affected how notes were represented to the networks. If the network determined during training that "one" in the soprano was most often harmonized with "one" or "eight" (the dominant) in the bass, the network shouldn't have had the option of averaging to find an output of "four" or "five."

For this reason, all twelve half steps of the chromatic scale (within a given octave) were represented by separate nodes in both the input layer and the output layer.

In the hope that the networks would do more than just find the most probable output note for any given input note, three (or sometimes only two) notes of the soprano line were inputted to the network simultaneously for each cycle. On each cycle, the network was provided with a past note, a current note, and a future note. Thus, the network was actually making an association between the correct bass note (during training) and a *pattern* of soprano notes. The goal was that the harmonization of the soprano note at a given state would depend in part on the past and future soprano notes.

In order to facilitate this idea, the input layer was actually composed of thirty-six nodes. The past note was always represented as a node between one and twelve, the current note between thirteen and twenty-four, and the future note between twenty-five and thirty-six. At the beginning and end of each chorale, of course, only two input nodes were activated, since there is no past note before the first note of the piece, and no future note after the last note of the piece. Figure 1 summarizes how notes were represented as input.

Table 1
Input representation

Letter name for note in the key of D	Degree of note on chromatic scale, starting with D as 1	Input line given to network including past and future notes
D	1	13,25
D	1	1,13,36
C#	12	1,24,34
B	10	12,22,32
A	8	10,20,25

It is important to note that rhythm information was almost entirely discarded when notes were inputted into the networks. There are many problems with representing rhythm in the input, and the information was discarded mostly for the sake of simplicity. The problem was handled by only treating a time step as relevant if a note during that time step changed, or was articulated, in either the soprano or bass line. In other words, if both the soprano and bass line held the same note for four beats, or an eighth of a beat, it was represented to the network as one input group. If, however, the bass line held a half note, while the soprano line changed from one quarter note to another (whether or not the actual pitch changed), this was represented as two input groups with the same bass note. The opposite is true also, in that if the bass note changed while the soprano line held a note, this was represented in multiple input groups. This method of extracting pitch changes while discarding rhythm is illustrated in Figure 2.

Table 2.
Extracting pitch changes and discarding rhythm

Letter name for soprano note in the key of D	Letter name for bass note in the key of D	Input line	Degree of bass note, given for training
F#	F#	3,17,30	5
G	F#	5,18,32	5
A	F#	6,20,30	5
G	B	8,18,30	10
G	C#	6,18,29	12
F#	D	6,17,27	1
E	C#	5,15,29	12
F#	C#	3,17,30	12

3. Network Architectures

Three network architectures were used. All the networks were created in Tlearn, a free program available for download on the internet. All the networks had the same input and output nodes, but differed in their middle layers. Each network had thirty-six input nodes. Each cycle (or “sweep,” as they’re referred to in Tlearn) had either two or three notes of input, and each was represented by activating the appropriate node in the input layer. Similarly, all the networks had twelve output nodes. In all three networks, every input node was connected to every output node. The weights on these connections were initially set to random numbers, but once the random values for the weights were determined, they were kept consistent every time the networks were reset. Figure 3 shows the input and output layers of all three networks. The arrows from the input layer to the output layer represent connections from every input node to every output node. Keep in mind that the network shown in Figure 3 is simply an illustration of the nodes and connections that all three of my networks had in common, and that the exact network shown in Figure 3 was not actually used.

3.1 Recurrent network

The first network I used was a recurrent network, in that its output from one cycle was used as input in the next cycle. This was done via twelve “copy-back” nodes, which simply took the exact activation values of the output nodes in one cycle and sent them as input for the next cycle. Recurrence seemed like a good idea for a harmonizing network because Bach’s bass-lines often move in what is known as *stepwise motion*. Stepwise motion is simply the musical term for music that moves along a scale one step at a time, without skipping directly from a low note to a high note, or vice versa. By allowing the network to be aware of its output on the previous cycle, the hope was that it would learn to associate one cycle’s output with a new output that was only one step away from the previous output (assuming this is what Bach most often did in this situation).

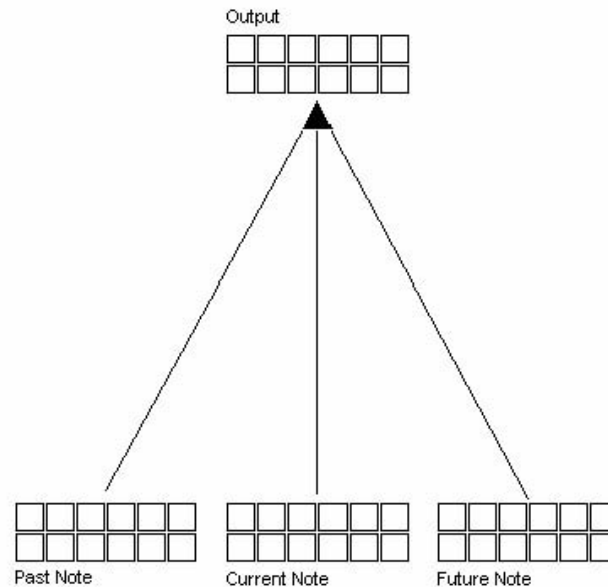


Figure 3. Nodes and connections held in common by all three networks used

Another goal of the recurrent network was to learn *resolution*. In Bach's pieces, and most traditional music, certain notes are extremely good predictors of the next note. For example, when chromatic scale degree twelve (the subtonic) moves to degree one (the tonic), it is said to "resolve." This resolution from scale degree twelve to one is very powerful, and listeners (those used to Western music at least) can predict it even with no knowledge of musical composition. At the very least, I hoped the recurrent network would learn this common resolution. A diagram of the recurrent network can be seen in Figure 4.

3.2 Pre-wired network

Another technique used in creating these networks was pre-wiring. In this network, I created a layer of twenty-one hidden nodes, seven for each set of twelve input nodes. Each node in the hidden layer was connected from three of the twelve input nodes directly below it. This is shown in Figure 5, but only three of the connections are shown so that the diagram is understandable. Each of the seven nodes in each of the three sections of the hidden layer represents a major scale degree. Unlike the chromatic scale, the major scale has only seven notes, and each of these notes has a chord associated with it. In a major piece, the chords associated with the first through seventh major scale degrees are major, minor, minor, major, major, minor, and diminished, respectively. So the first node in the hidden layer is connected from input nodes one, five, and eight, composing the major chord associated with scale degree one. The second hidden node is connected from input nodes three, six, and ten, and so on.

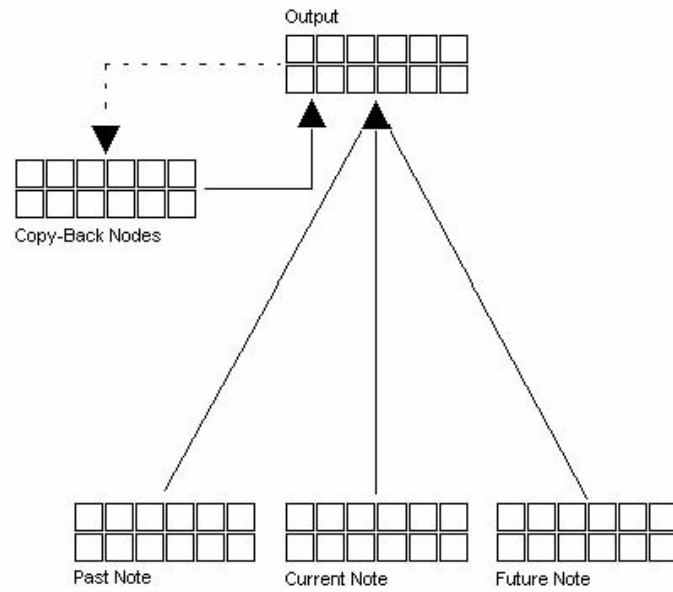


Figure 4. The recurrent network

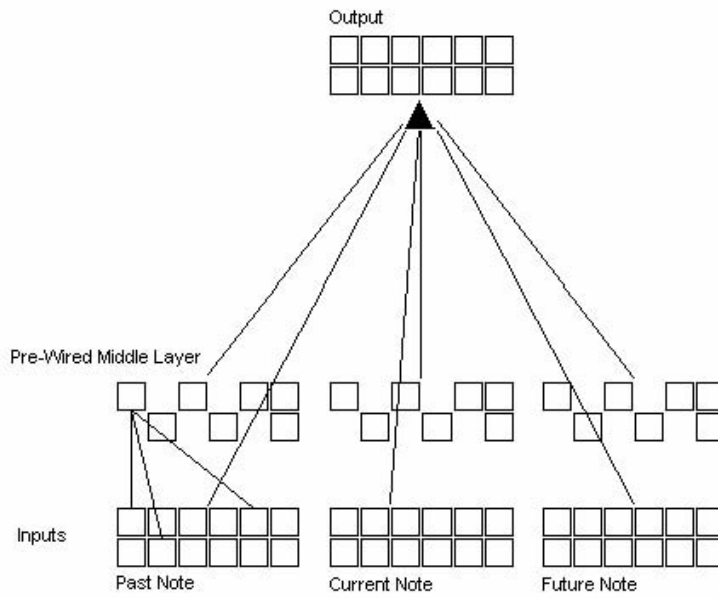


Figure 5. The pre-wired network

I chose to pre-wire a network for two reasons. First, I simply hoped that the network would be able to make associations between notes of the same chord within the key, since this could conceivably make the network a better harmonizer. Secondly, since I was, in part, attempting to model the human ability to harmonize Bach chorales, I thought it only fair that my network be explicitly taught some characteristics of music the way most humans are when they're learning to harmonize music.

3.3 Recurrent and pre-wired networks

This third network is simply a combination of the previous two. I created it in the hope that it would outperform both the recurrent and pre-wired networks, by using both techniques to find the best harmonization. This network is represented in Figure 6.

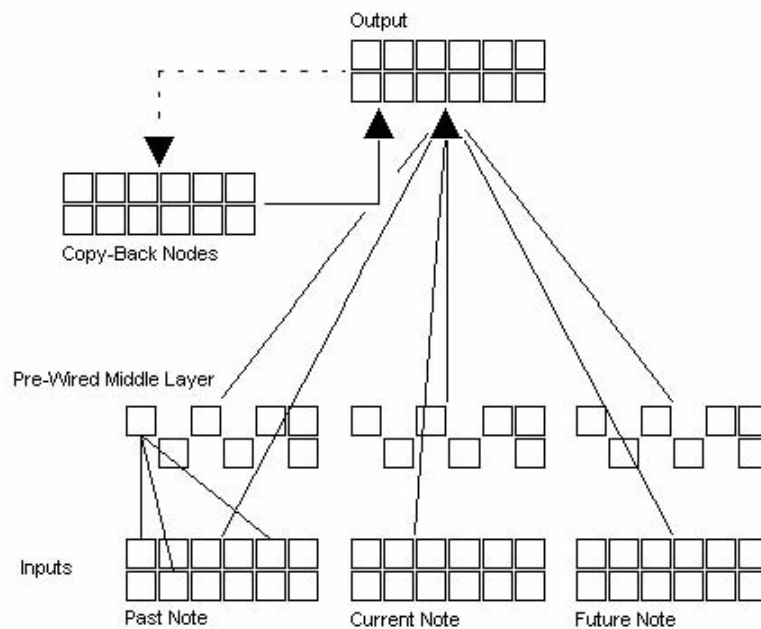


Figure 6. The recurrent and pre-wired network

4. Training and Output Representation: Numbers to Notes

The networks were trained on a total of seventeen chorales. Chorales were not used if they were in a minor key, or if I felt certain that there was a key change within the piece. The seventeen chorales provided the networks with a total of 839 lines of input, however I also stopped and tested the networks' performance after 471 lines of input, so that I could measure the networks' improvement as input was added. Before testing, the networks were all trained on five-thousand sweeps, and the learning rate was set to 0.1. I tested the networks on two pieces that I felt were indicative of Bach's harmonization style. Their titles are "Open wide for me the portal" and "My trust is bold." When tested, the networks provided me with the activations of the output nodes for each input line of testing. I chose to use the maximum activation of the output layer for each input line as the networks' "best

guess” as to the correct bass note. After determining the maximally activated output node for each input line, I regarded the number of that node as the chromatic scale degree for the output note. I used Finale (music software) to place these notes onto a staff, conserving Bach’s soprano line, and his rhythm and octave distinctions as much as possible. Finale also allowed me to make audio recordings of these pieces as MIDI files.

5. Tests of Success: Percent Accuracy and Average Step Size

I developed two methods for testing the success of my networks’ outputs. My first approach, percent accuracy, seemed like the most straight-forward method. I simply compared every output value to Bach’s original bass line. For each piece, I divided the number of outputs that matched Bach’s to the total number of outputs. Figure 7 shows the percent accuracy of each network after 839 lines of input. The percent accuracy is averaged over the networks’ output for both “Open wide for me the portal” and “My trust is bold.” As seen in Figure 7, the network that was both pre-wired and recurrent outperformed the other two networks, as predicted.

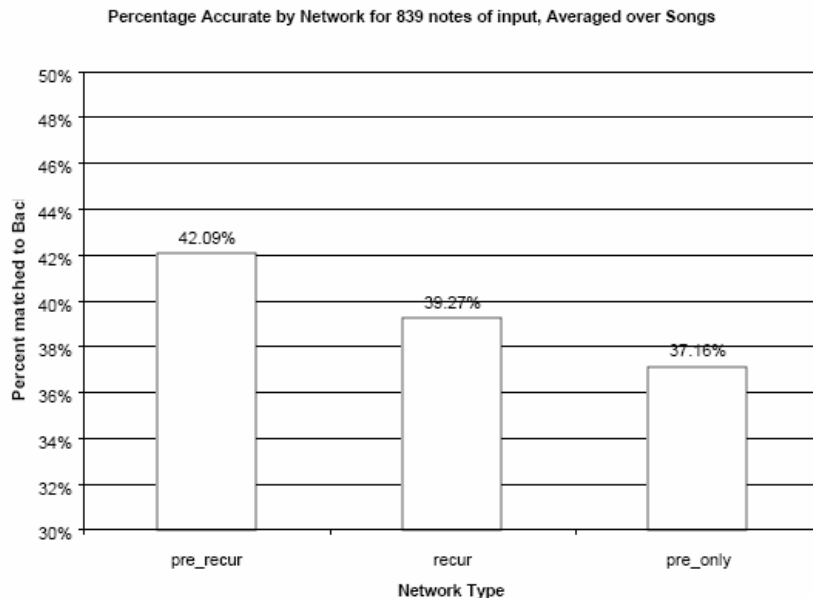


Figure 7. Percent accuracy for each network after 839 lines of input

It should be noted, however, that there may have been a significant problem with using percent accuracy as a measurement of the output’s “goodness.” The percent accuracy of a piece was determined independently of any given note’s duration or position within the piece. So although two outputs may have the same percent accuracy, as determined here, one of the outputs may actually sound much better because the notes that were correctly guessed are longer and in crucial places within the piece. A potential solution would be measuring percent accuracy for time steps, so that each half beat, for example, would be equally

weighted in the calculation of percent accuracy. Furthermore, certain time steps could be somehow weighted as more important than others based on their position in the song. For example, the first and last notes of a piece often play a large role in how “good” the piece sounds to a human listener (this is at least in part due to primacy and recency effects) and could therefore be weighted as more important than other notes when determining percent accuracy.

The second method of measuring an output’s “goodness” was the average step size between notes in the output. As mentioned previously, traditional compositions such as Bach’s often include a lot of stepwise motion. As a result, the average step size between notes is very small. In fact, the average step size of the bass line in both Bach’s “Open wide for me the portal” and his “My trust is bold” is only about 2.4 (half steps). Figure 8 shows the method I used to determine average step sizes for both Bach’s pieces and the output of my networks. Figure 9 shows the average step sizes for Bach’s pieces as compared to my three networks. Contrary to my predictions, the pre-wired network had smaller average step sizes in its output than the recurrent network. However, another goal of the recurrent network was to see resolutions from chromatic scale degree twelve to one, and this does occur in the recurrent network’s output more than the output of the other two networks.

<p>IF $x_i - x_{i+1} \leq 6$</p> <p>THEN $s = x_i - x_{i+1}$</p> <p>ELSE $s = 12 - x_i - x_{i+1}$</p> <p>Where s equals the step size (not the average step size), x is equal to the chromatic scale degree of the note, and i is equal to the serial position of the note within the chorale.</p> <p>And now $S = \frac{s}{n-1}$</p>
--

Figure 8. Determining average step size

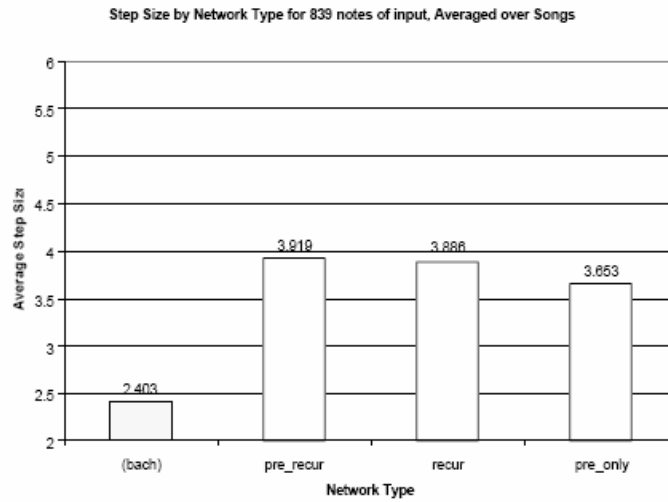


Figure 9. Average step sizes for Bach and all three networks after 839 notes of input, also averaged between both test songs

For further analysis, I've included graphs of the percent accuracy for the song "Open wide for me the portal" in Figure 10. This figure shows the percent accuracy for all three networks, after both 471 lines of input, and after 839 lines. Although the network that was both recurrent and pre-wired improved with more input, the other two networks had higher percents of notes matched to Bach's after only 471 lines of input. One possible explanation for this is that the chorales entered after 471 lines of input had more erratic bass lines that were not as easy for the recurrent networks to assimilate. Figure 11 shows the average step size for the same test piece, and as expected the average step size decreased with more input.

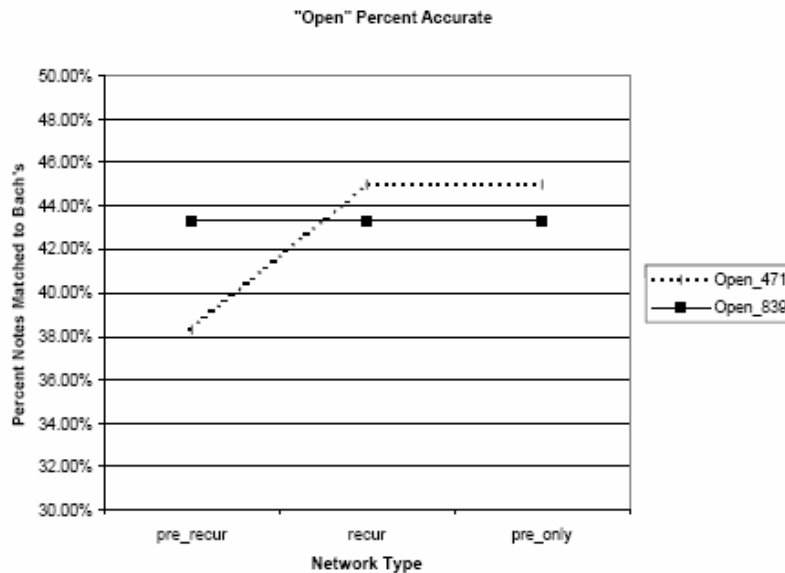


Figure 10. Percents accurate for "Open wide for me the portal."

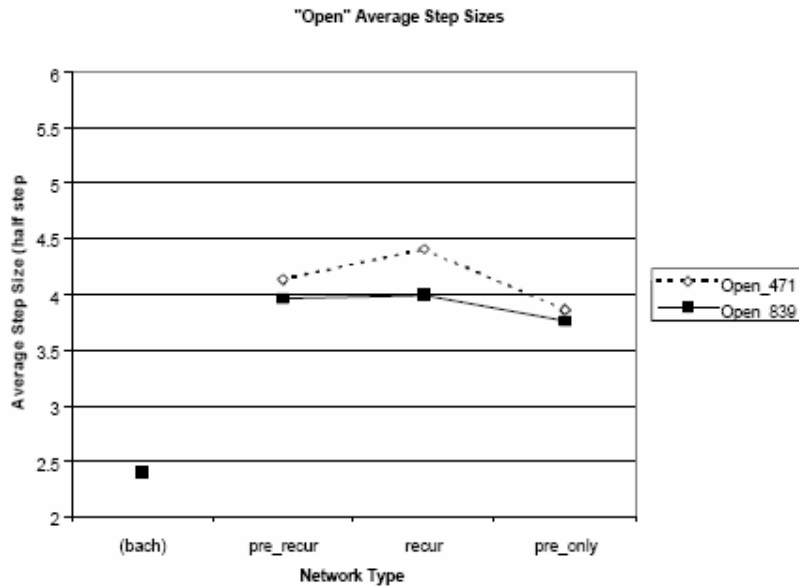


Figure 11. Average step sizes for “Open wide for me the portal.”

Figures 12 and 13 are similar to Figures 10 and 11. They present the same information, respectively, for the chorale “My trust is bold.” Once again the average step size decreases with more input, as seen in Figure 13. Figure 12 shows that the percent accuracy improves with more input, except for the recurrent network. This fact further supports the possibility that the later chorales inputted to the networks were less appropriate for the recurrent network, and more suited for the pre-wired networks.

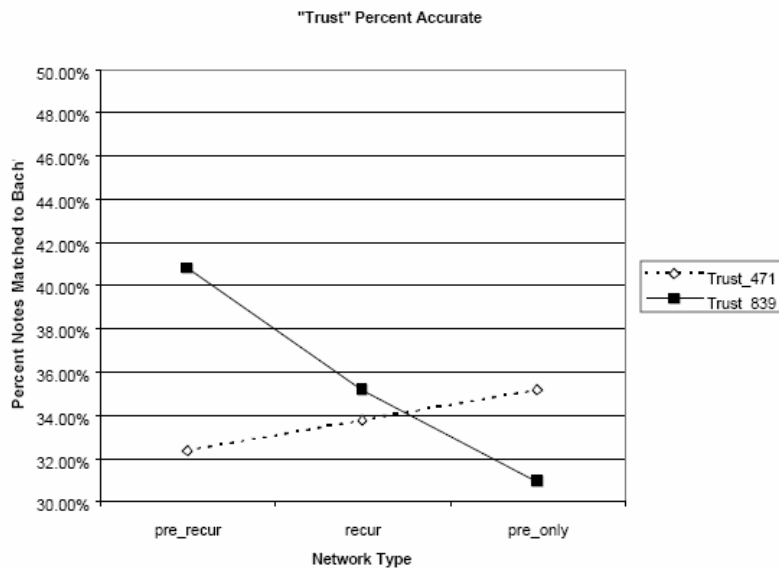


Figure 12. Percents accurate for “My trust is bold.”

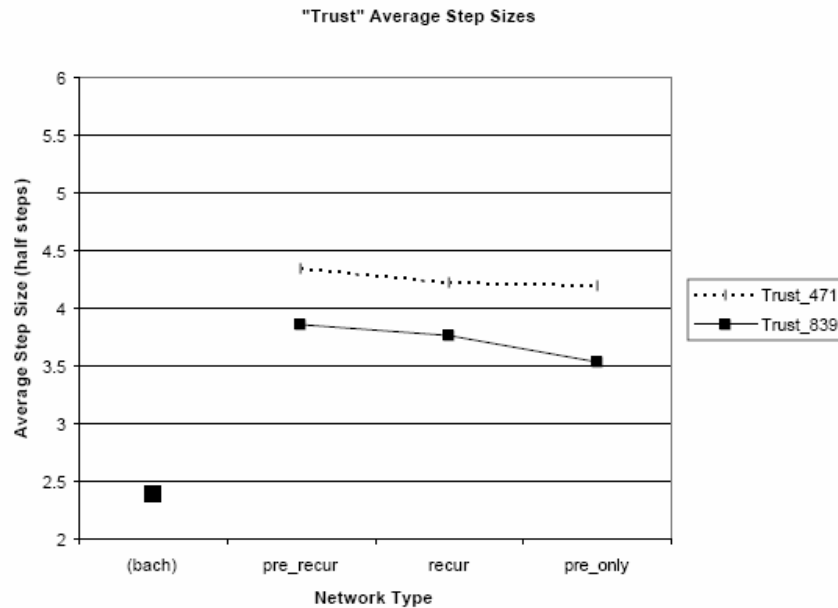


Figure 13. Average step sizes for “My trust is bold.”

6. Problems and Possible Solutions

Artificial neural networks find the most probable outputs by strengthening associations between inputs and correct outputs during training. The network’s ability to propose “creative” output is significantly hindered, as it is, in essence, designed to do the exact opposite. Thus, my networks never chose notes other than major scale degrees as output. Bach, on the other hand, occasionally used accidentals (notes that are not major scale degrees) to add to the creativity of his pieces. A possible solution to this would be to have occasionally insert some randomness into the network’s output, however this would probably end result in some unpleasing outputs. Another possibility that might provide more creative outputs would be to use a loose choice rule to determine the output for each input line, as opposed to the maximum activation rule I chose to use. This way a little randomness is used, but the likeliness of an output node being chosen is weighted by its activation. This could provide a nice output, since output that the network is very confident of would most likely not change, but when activation is a little more evenly dispersed throughout the output layer, more interesting choices may be made.

Another problem with the networks’ outputs was the inability to handle runs in the soprano and bass. A “run” in music is usually a fast-paced (eighth notes or shorter) section of a melody or harmony, and usually contains a lot of stepwise motion. Because my networks only had information about one note in the past or future at most, they were unable to model composition of runs in the bass. Again, the result is a less “artistic” harmonization. In the future, it would be interesting to let the network see several steps into the past and future of the soprano line, and at least two steps into the past of the bass line. In the hopes of modeling

human harmonization, it seems only fair to provide the network with this information, since obviously any human attempting to complete the same task would have such knowledge.

Another problem already mentioned was the difficulty with representing rhythm to the network. A way of fixing this problem might be to add nodes to the input layer that would represent the duration of each note. So along with representing the pitch of each note in the input, another one of five nodes (perhaps) would be activated that would tell whether the note was a fermata or whole note, half note, quarter note, eighth note, or shorter. The problem with this solution, however, is that it does not allow the bass line to move freely of the soprano line. The ability for one soprano note to be harmonized with several moving bass notes is one of the most important parts of harmonization in Bach's style, so this solution would have to be altered if not entirely abandoned.

Despite the problems with my networks, I am pleased with the results. The fact that their output approached fifty-percent accuracy to Bach's notes is promising, to say the least. In the future, I hope to improve both the networks' percent accuracy and average step size by implementing some of the ideas I've mentioned here.

7. Acknowledgments

Special thanks to Erik Isaacson, Sean McLennan, Robert Goldstone of Indiana University Bloomington and the COGS Q400 students of Spring, 2001.

Spatial Presentation and Compatibility of Horizontally and Vertically Associated Words

Jeffrey Gilleland

Cognitive Science Program, Indiana University Bloomington

1. Introduction

When many people read books, analyze images, or interpret graphs, they do so from left to right and top to bottom. This is the way that most English books are written, the way the English language is written, and the way the English language is read. If this is true, then why has research shown that words in the right visual field have a higher recall and accuracy rate than words presented in the left visual field? This finding is very interesting and is further explored in the following research. The following experiment manipulates and explores the aspects of existing publications to further dissect and understand this bizarre phenomenon and produce more “fitting” results.

Previous work by Timothy R. Jordan and Geoffrey R. Patching shows words presented in the right visual field (RVF) yield a higher accuracy rate of recall than words presented in the left visual field (LVF). How might this come about? Words viewed in the RVF are done so with the left sides of each retina and words in the LVF are seen with the right sides of each retina. The information from the RVF is first sent to the left hemisphere of the brain and information from the LVF is first sent to the right hemisphere of the brain. The left hemisphere is responsible for language processing in 93% of all people (96% in those that are right handed). Words from the RVF thus have a direct pathway to the language processing part whereas words from the LVF must first go to the right hemisphere to be processed and then sent across where it can be further processed as language. This pathway is much noisier and so the message can be delayed, disrupted or forgotten, significantly reducing reaction time and/or accuracy for words presented in the LVF.

In the experiment by Jordan and Patching, a word was shown in either the RVF or in the LVF followed by the presentation of this word along with a distracter (in this case a different word than the previous). Subjects were instructed to select the correct word that had been shown. They found that words presented in the RVF had a higher accuracy rate than those presented in the LVF. In the following experiment words are presented in each of four equally divided quadrants to test for effects of positioning such as seen in the previously mentioned experiment. The following experiment not only explores the right and left visual fields, but also the upper and lower visual fields. This was not taken into consideration in their experiment. One aspect of this previously conducted experiment is that the words were only presented for a brief portion of a second. Here, six words will be presented for about 2 seconds, allowing for some search patterns to develop. This will not only test visual fields, but also the biases that may have been previously instilled through years of exposure to the

English language.

This experiment also explores the effects of presenting words vertically against horizontally. In doing so, the effects of word compatibility will be tested through the spatial presentation of the word. Most words have an implied direction or motion associated with them, either horizontal or vertical. For example, the word lift would be associated with a vertical connotation and the word push would have a horizontal motion associated with it. A word would be considered compatible if it was presented with the same spatial orientation as its implied associated direction. So the word lift presented vertically would be considered compatible, while the word push presented vertically would be considered incompatible. The hypothesis for this experiment is that words presented in a compatible manner will produce higher accuracy rates and recall than those presented in an incompatible way. Furthermore, this will be true for both the horizontally associated words and the vertically associated words.

Several other visual field biases are discussed in the previously described experiment as well as others, such a complimenting study conducted by Iain T. Darker and Timothy R. Jordan. Subjects in this experiment were shown a blinking focal point that they were told to focus on. After the point blinked rapidly for a short time, either a four-letter word or a four-letter non-word was presented to either side of the focal point. This was presented momentarily and the subject was asked to identify which word they had seen. This experiment not only tested for differences in the RVF and the LVF but also for differences in the upper and lower visual fields. Results from this experiment as well as others produce, at best, variable results for the upper and lower visual fields. However, with the layout of the following experiment, differences between these visual fields are expected. An upper visual field bias is expected for search patterns that will lead to higher accuracy rates and recall for words presented in the upper visual field. This bias again would be the result of schooling and reading techniques developed earlier on.

The Darker and Jordan experiment brings up yet another factor that must be considered when presenting words. Most words have what are referred to as “in-word” cues that can be used to identify a word at a glance. For example, the word “heat” has the word “eat” within it that would give someone glancing at the word an added clue for future recognition of the word. A paradigm that tries to eliminate these in-word cues is known as the Reicher-Wheeler task, which takes words that are very similar and differ only by one letter. This paradigm can be seen by using the words “heat” and “neat”. The in-word cues for both of these words are exactly the same, varying only by the first letter, which is known as the critical letter. Without knowing the critical letter of each word, there is no way to determine which word had actually been seen.

When given a chance to scan a list of words, people will tend to start in the upper left quadrant and then scan the rest of the list based on the way Americans have been taught to read. Therefore, a further prediction of this experiment is that words in the upper left quadrant will have the highest accuracy rates and recall, and the words in the lower right quadrant will have the lowest accuracy rates and recall. Additionally, words presented in a compatible manner will have higher accuracy rates and recall than words presented in an incompatible way. These predictions are based on the way reading and writing are taught, and given enough time to scan and analyze a word list, these engrained processes will come through in the form of search biases.

2. Method

2.1 Participants

Ten participants were used in this experiment. Five were females and five were males, all college students at Indiana University Bloomington. All but one of the participants was right-handed, although this should not affect the results in any way.

2.2 Materials and Design

There were different types of words used in this experiment, including vertically associated (VA) words, horizontally associated (HA) as well as some relatively neutral words. Some of the VA words included lift, fall, jump, and rise. Some of the HA words included roll, path, road, and kick. Some neutral words would be moth, drug, and cake. All the words used in the experiment were four letters in length and had a familiarity and concreteness rating of at least 350, as defined by the MRC Psycholinguistic Database. This means the words used in the experiment were relatively common words and recognizable by the subjects.

Six words were presented either horizontally or vertically on a slide in one of four quadrants. The words were presented either in a compatible way or in their incompatible way, according to the detailed description mentioned earlier. The words on the slide had a font size of 24 and were presented in all caps so they were easily noticeable and easy to read.

The subject was instructed that the slide was going to be shown for a brief amount of time and to remember as many of the words as they could. The slide was then presented to the subject for two seconds and then a blank white screen was shown. The subject was then instructed to recall and write down as many of the words as they could remember. The next slide was then shown with the next six words and the same process was repeated. The subject was shown one practice slide to demonstrate what the slide might look like, in order to give them adequate preparation to achieve accurate results. The subject was shown four slides containing vertical words and four slides containing horizontal words. This gave a total of 24 vertical words and 24 horizontal words, all presented in different quadrants and with different compatibilities.

The subject was judged based on the accuracy of the words they could recall. Accuracy was determined based on different levels of analyzation. Their scores were analyzed based on percent correct for each quadrant and percent correct based on the compatibility of the word.

3. Results and Discussion

Table 1 shows the amount of correct responses for all of the slides combined. It is divided by quadrant and also by compatibility. There were a total of 16 words displayed in the upper left quadrant, 15 in the upper right, 13 in the lower left, and 12 in the lower right. There were a total of 15 compatible words and 14 incompatible words displayed. These numbers were also converted into percentages as seen in Figure 1.

Table 1.
Number of correct responses by compatibility and quadrant

Subject #	Quadrant				Compatibility	
	Upper Left	Upper Right	Lower Left	Lower Right	Comp.	Incomp.
1	15	2	10	3	11	6
2	14	9	9	5	13	7
3	15	2	6	4	9	7
4	11	6	6	3	6	7
5	16	9	7	1	13	8
6	14	9	4	4	11	8
7	13	6	8	4	12	8
8	11	8	5	8	10	6
9	15	9	11	0	12	8
10	14	9	7	4	11	9
Totals	138	69	73	36	108	74
Tot Possible	160	150	130	120	150	140
% Correct	86.25%	46.00%	56.15%	30.00%	72.00%	52.86%

Figure 1 depicts percent correct data for the different quadrants. Words located in the upper left quadrant were correctly recalled 86.25% of the time, words in the upper right 46% of the time, words in the lower left 56.15% of the time, and words in the lower right 30% of the time. The standard error for this examination was 5.0, which means the percent recalled in each quadrant were significantly different from each other. The overwhelming percentage here is 86.25% for the upper left quadrant. This is followed by the next highest percentage, the lower left quadrant. This means the subjects tended to start on the left side of the screen (particularly in the upper left) and then worked their way down and across the screen in search for words. With only 2 seconds available for search and store, it was very difficult for any of the subjects to recall more than 4 of the words from any one slide. There were a few subjects that recalled 5 words on a given slide but no one subject was able to recall all six words from any given slide.

This was done intentionally in order to test for search patterns. With inadequate time for the subject to scan the entire screen, search patterns were developed and biases were seen. The first place that almost all the subjects looked was to the upper left and then either down or across. When recalling the words, the first word subjects recalled would often times be the word from the upper left quadrant.

The next area that was tested was the compatibility of the word. Subjects recalled words presented compatibly 72% with a SEM of 4.32, while only recalling the incompatible words 53% of the time with a SEM of 2.31. This indicates words that were presented in their compatible fashion yielded a significantly higher rate of recall than words presented in an incompatible manner. This can be seen in Figure 2. This figure is the visual representation of the percentage correctly recalled for each type of word.

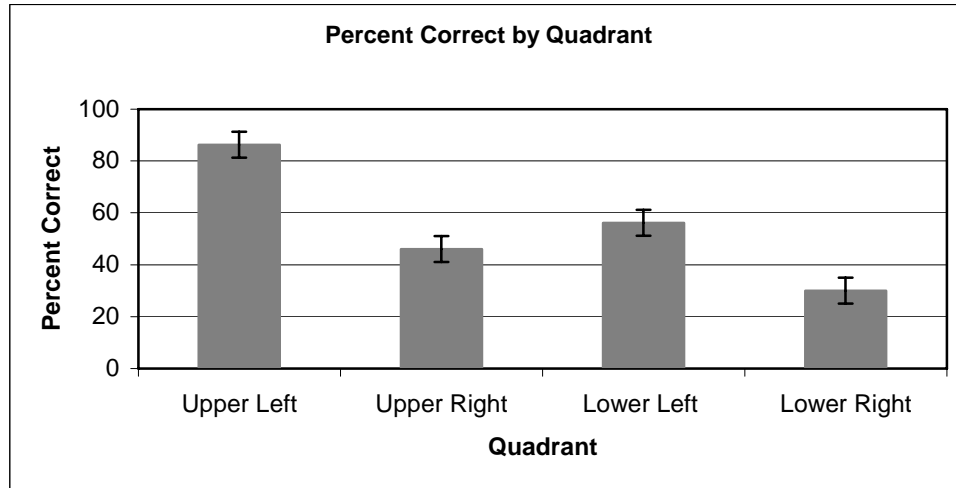


Figure 1. Percent correct responses by quadrant

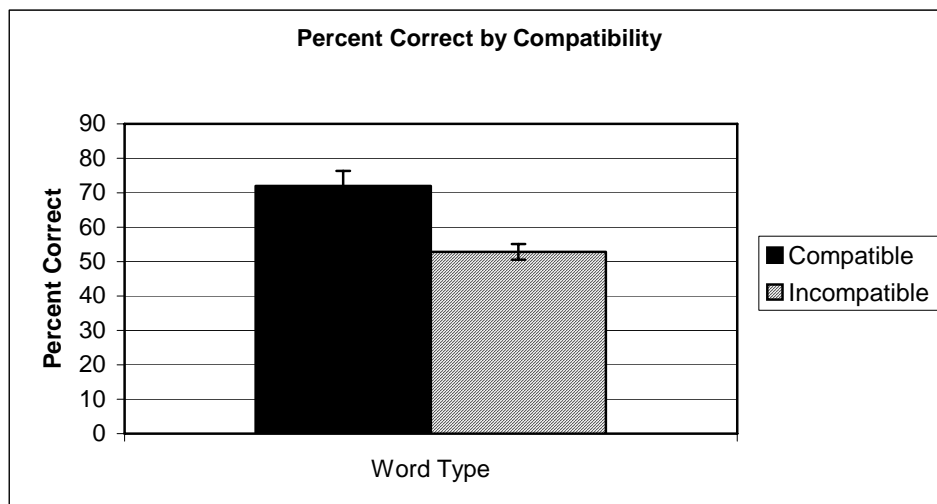


Figure 2. Percent correct responses by compatibility

This data supports the hypothesis that the upper left quadrant would have the highest rates of recall and the lower right quadrant would have the lowest rates of recall. The data also shows us that the lower left quadrant had significantly higher recall than the upper right quadrant, which means the overall search pattern was from left to right.

Furthermore, the data provides support to the hypothesis that the compatibility of a word will play a difference in storing and recalling. The compatible words, whether they were presented horizontally or vertically, had significantly higher rates of recall than did incompatible words.

4. Conclusion

This experiment was able to show that when given enough time to develop any type of search pattern, when analyzing words for recall, the upper left quadrant will produce the highest rates of recall. This is most likely due to the way people have been taught to read and analyze problems. Given time to search and analyze, the visual field biases that were seen in the previous experiments were essentially eliminated by the search patterns of the subjects.

In addition, the compatibility of the word played a significant role in recalling task for the words. This may have been due to people associating the words with their implied movement. It was easier for people to associate words when they were in their compatible state. Words presented in a normal way (horizontally) will still be the most familiar presentation of a word, but when only comparing compatibility and not taking this aspect into consideration, it is clear that compatibility also plays a role.

References

- Darker, Iain T., Timothy R. Jordan. (2004) Perception of words and non-words in the upper and lower visual fields. *Brain and Language*. 593-600.
- Jordan, Timothy R., Geoffrey R. Patching. (2002) Assessing effects of stimulus orientation and perception of lateralized words and non-words. *Neuropsychologia*. 1693 – 1702.

A Hybrid Neural Network/Finite-State-Machine Model of Adversarial Artificial Intelligence

Daniel S. McFarlin

*Department of Computer Science; Cognitive Science Program
Indiana University Bloomington*

1. Introduction

Adversarial Artificial Intelligence (AAI), the simulation of human-like opponents in a domain specific context, has been a cornerstone of AI research since the inception of the field (Cheong, 2005). One of the original aspirations of the field was to create a chess playing program that could routinely trounce grandmasters. The original “Proposal For the Dartmouth Summer Research Project On Artificial Intelligence” contained such a goal (McCarthy et. al, 1955). Algorithmically, devising such a chess playing AI proved to be tractable. The initial component, the minimax algorithm, was developed and proven by John Von Neumann in the mid 40’s (Von Neumann, 1944). Combined with alpha-beta pruning, developed by Claude Shannon, chess playing programs have gone on to challenge and best the world’s finest (Schaffer, 2001). Achieving that goal in practice, however, had to wait nearly 50 years for computational resources to mature to the point where the hardware could sustain and process the massive search spaces required to win this two-player, zero-sum, transparent game (Schaffer, 2001).

Modern AAI theoreticians and implementers long for the days of minimax, alpha-beta pruning and (relatively) straightforward games like chess. The modern electronic adversary now inhabits a rich, 3-D world where the gamestate space dwarfs that of the humble chessboard. Where there were once two players (one electronic, one human) there are now an arbitrary number of human and electronic opponents. Added to this is increasing demand for realistic, innovative and doctrinally-sound adversaries. Note that realism in modern AAI may include “artificial stupidity” as one of its distinguishing characteristics (Linden, 2001). This is a far cry from the near-perfect chess-playing automata of old. However, just as the original, algorithmically-sound chess playing programs were hamstrung by a lack of computational resources (both processing speed and memory storage capacity), so too is modern AAI in its most bleeding-edge incarnations: military simulators and first-person-shooter (FPS) video games. These time-constrained applications must (by virtue of their design goals and intended purposes) feature realistic AAI that now must compete with a host of other processes and functions for vital CPU time and memory storage (Carmack, 2005). This paper examines the approaches to modern AAI in these applications from two perspectives: the AAI theoretician and the game/simulator AAI implementer and proposes a hybrid system incorporated into a classic 2D FPS: Wolfenstein 3D.

2. The Domain:

2.1 *A primer on the domain*

Before examining the approaches to AAI in academia and games it is useful to understand the problem domain. One of the most challenging AAI domains and the focus of this paper is squad maneuver warfare. Squad maneuver warfare involves military combat operations of the smallest, organized, and singly-command unit called the squad. Squads typically contain around one dozen elements be they humans, tanks, boats etc. Squads are further subdivided into fireteams. The classic example (and one that has heavily influenced simulator and game design) is the US Marine Corp squad (McBreen, 2001). This squad consists of three four-man fireteams all of which are collectively commanded by a single individual. The elements of this squad are not homogenous in terms of capabilities. Obvious differences are cognitive ability but less obvious are offensive and defensive capabilities. A Marine fireteam generally consists of three rifleman and one heavy-weapon element. A heavy-weapon can include an automatic rifle, a machine gun or a guided-missile. Squads are generally assigned specific objectives with the task of devising a plan to accomplish the goal left to the squad. An objective might include defending a section of territory against an enemy attack for a particular amount of time. The squad generally operates in accordance with a military doctrine consisting of standard operating procedures involving tactics, strategy and logistics. A military doctrine is effectively a description of the characteristic behavior of military units with accompanying heuristics designed to achieve these characteristics. The degree to which doctrinal adherence is enforced and emphasized varies from organization to organization.

Doctrinal adherence directly impacts perceived realism. There is a notion that a military unit “fights like Afghans”, or “fights like Germans”, etc. US forces are unlikely to employ “human-wave” attacks even though such a tactic may grant tactical surprise and be of comparable utility to that of a doctrinally-endorsed pincer-movement style assault. In the former, soldiers rush an enemy defensive position in a line-abreast (think offensive-line formations in football) formation, attempting to focus as much firepower on the enemy position as possible while relying on speed, shock and surprise to overwhelm the defenders. In the latter, one fireteam (the suppressing fireteam) focuses its fire on the defender’s position to confuse and distract while remaining fireteams advance towards the enemy position from the extreme left and the extreme right. The fireteams on the periphery are assaulting the defender’s position and defending the suppressing fireteam simultaneously. Implementing this with people is difficult. Implementing it computationally is even more difficult. To date, it has only been done in a restricted form in the most recent (October 2005) games and simulators.

2.2 *Broader Implications of the Domain:*

One may reasonably question the broader utility of this domain in illuminating AAI as a whole. Squad maneuver warfare encompasses a great many cognitive activities: perception, complex pathfinding over varied and dynamic terrain, time-constrained planning and decision making involving large numbers of multi-attribute factors, inter-agent synchronization and communication just to name a few. Although general-purpose AAI would be preferable its attainability is much in doubt given the fragility, complexity and lack of scalability of some of the more generic approaches (Pew, 1998). In contrast, research into the elements of cognition responsible for simple motor abilities in primates has shown great

specificity and isolation of neural hardware both for input and output in accomplishing motor movement (Schieber and Hibbard, 1993). Similar results have been shown for other cognitive tasks as well (Churchland, 1994). Consequently, the prevailing intuition for squad maneuver warfare is that the problem domain should be tackled with a computational architecture of highly specialized and task specific components (Liden, 2002).

Finally, it has been observed (and in some cases implemented as such) that an individual AAI is a squad unto itself in terms of pathfinding, time-constrained planning and the other cognitive activities (Orkin, 2003). The individual AAI must coordinate various sub-processing mechanisms much as a squad must coordinate its actions. In fact, expectations of realistic behavior are generally higher for individual AAI than for squads. Squads generally exhibit less flexibility than an individual in both real and artificial world environments. Overall, squad or individual maneuver warfare simulators/games pose significant challenges for AAI theoreticians and implementers.

The challenge for the modern AAI in military simulators/FPS games can be summed up from the perspective that academic models are too general to be specifically useful whereas game implementer models are too specific to be generally useful (Isla, 2005). There is even some debate in the upper echelons of the industry if substantial investment in next-generation (read: academically developed models) AAI is worth the effort. John Carmack, the legendary game designer who developed the engines for Wolfenstein 3D, the Doom franchise and the Quake franchise has recently stated his view that AAI is mostly “a matter of scripting” and that further investment in more advanced forms of AAI is generally a “waste of time” (Carmack, 2005).

It should be noted that the majority of top-tier AAI implementers vehemently disagree (Tozour, 2005). The author would also like to indicate that Carmack’s latest effort, Doom 3, was much lampooned for its laughably bad AAI. Despite Carmack’s statements to the contrary, the significant utility of advanced AAI models is generally agreed upon.

3. Theoretical Vs. Applied Models:

Connectionist and production-rules models developed in academia are of enormous theoretical and practical importance in many areas of cognitive modeling and simulation. Within the context of the academy, such models are sufficiently fast (though not generally run in real-time) and often constrained to very specific subdomains to ensure that they are computationally feasible and sufficiently easy to analyze and explain (Hoffman, 2005). The explanatory powers of these models and their connection to human cognition (presumably the basis for the decision-making capabilities of an electronic adversary) would seem to suggest them heavily for roles as AAI in games and military simulations. Indeed, one of the more recent and well received games, F.E.A.R., utilized a hybrid production system modeled on SOAR and a planning definition language based on FDDL (Laird, 2004). The architecture is quite revolutionary and the game has received many accolades. Fundamentally, the architecture consists of a blackboard containing working memory elements (Orkin, 2005). These working memory elements are used to satisfy preconditions for a series of behaviors. The planner then searches for a sequence of behaviors. This sequence forms a plan which is used to accomplish a goal. The goals and behaviors are specified by the game designer at compile time.

The F.E.A.R. system is inspired by the STRIPS architecture and shares many similarities (Nilsson, 1998). One similarity is the presence of a suite of standalone sensors that deposit input into working memory. These sensor arrays are both interrupt and poll driven. This hybrid sensor suite design was necessitated by computational limitations; many of the pathfinding algorithms involve raycasting which would be prohibitive to perform on demand. Consequently, raycasting is performed periodically and the results taken into account only when they are needed. In contrast, an interrupt-driven sensor would be one responsible for detecting dangerous foreign objects such as a grenade. If the sensor detects a disturbance (in the form of grenade perhaps) it immediately inserts a new working memory fact onto the blackboard (`kSymbol_DisturbanceExists`) with a value of true assigned to the symbol. The symbol also has metadata associated with it, called attributes, such as 3D position, direction and other properties such as object type.

Each attribute has a confidence value associate with it. Here, confidence is used to indicate the amount of “noise” associated with an attributes value. For example, the confidence value (0.0 – 1.0) of a disturbance’s 3D position is influenced by a probabilistic model that takes into account the environment in which the disturbance was detected, the auditory and visuospatial abilities of the agent detecting the disturbance and the current goal state (hunting an enemy, patrolling, etc) of the agent. Interestingly, every working memory fact contains a desire attribute whose confidence value is used to denote the agent’s interest in acknowledging the fact. This value is influenced by the amount of relevant working memory on the blackboard. This quantity can be thought of as dominating an agent’s decision making process and encouraging the agent to disregard new input data from the sensor suite. If the desire value is high, the planner may discontinue the current plan and generate a new plan which has the goal of setting the value of `kSymbol_DisturbanceExists` to false.

The planner attempts to generate a sequence of actions that satisfy this goal. Employing A*, the planner examines the set of possible actions whose preconditions are satisfied, whose total heuristic cost is minimized and whose effect is to satisfy the goal. Actions have associated heuristic costs which are specified by the game designer. For example, the action `LookAtDisturbance` has a lower cost than `InspectDisturbance` as the latter action generally poses less danger to the agent and entails fewer preconditions. Conversely, the `Attack` and `AttackFromCover` actions both satisfy the goal of setting `TargetIsDead` to true. Even though `AttackFromCover` entails more preconditions (effectively `MoveToCover` and `UseObject(cover)`), its cost is less than that of `Attack`. The intention of this design is to encourage use of the `AttackFromCover` behavior which is a more doctrinally sound approach. Additionally, actions may have context-specific precondition validation code that is invoked when the action is considered by the planner. These functions may remove the action from consideration depending on the sensory input values on the blackboard. For example, `LookAtDisturbance`, `InspectDisturbance`, `ReactToDanger` and `EscapeDanger` all set the value of `kSymbol_DisturbanceExists` to false. However, the first two actions specify that they are to be ignored in the presence of a dangerous object (such as a grenade) of high confidence value introduced by the disturbance. These precondition validation functions can be thought of as instincts. The selected actions may also contain post-processing functions which are invoked after the action is performed.

A nice feature of the blackboard approach is that a new plan does not overwrite the existing goal state and current plan. Rather, once the current plan has been executed the agent resumes what it was doing, possibly influenced by the result of accomplishing the just executed plan. For example, evading an exploding grenade means that dust and debris now fill the air which may cause the agent's confidence in the player's location to erode. Squad maneuver behavior is accomplished by giving agents access to the blackboards of the other elements in the squad. This transparency is combined with a reservation system for actions, positions and firing responsibilities or effectively any action in the system. For a squad, the number of slots in the reservation system depends on the size of the squad. A four man squad may have two slots designated for firing from cover, one slot designated for advancing and one slot designated for grenade throwing. This constraint architecture prevents all elements of the squad from engaging in the same action at the same time. Though simple in design, the squad maneuver system produces quite dynamic and doctrinally sound squad emergent behavior. The insight was to enhance individual squad members and to facilitate emergent squad behavior through simple coordination.

The F.E.A.R. approach draws upon many established models of decision-making and planning found in academia and eclipses the traditional approaches in game AAI such as hierarchal finite-state-machines and static rules systems. The author views this approach as a mix of SOAR and Decision Field Theory (DFT). Though not implemented using a connectionist model, F.E.A.R. incorporates many of the concepts expressed in DFT; in particular, the manner in which the confidence values fluctuate over time in relation to the external environment and the degree to which internal motivation ("desire" in F.E.A.R. parlance) influences decision making (Busemeyer, 2001). Though the F.E.A.R. approach has some advantages over the connectionist approaches, namely computational efficiency and transparency, its extensibility and maintainability are dubious at best. Though it may be a reasonable approximation of some level of human cognition in a specific sub-domain, no learning and only a limited degree of tactical innovation are possible given the fixed rule set. Lack of learning and limited improvisation hobbles the F.E.A.R. to a degree not seen in most connectionist models and even production systems such as SOAR. If F.E.A.R. represents what can be accomplished with a probabilistic production system the next logical step is a hybrid connectionist-rule system.

4. Wolfenstein 3D:

The approach implemented by the author combines a rule system in the form of a probability-based finite-state-machine with a neural-network that determines state transitions. The source code is available on request. Understanding the architecture first requires some understanding of the game. Wolfenstein 3D (Wolf3D) is a 2-dimensional (the illusion of 3D objects is provided by aspect permutations of 2D images called sprites in comparison to true 3D models in games like F.E.A.R.) FPS released in 1992 for MS-DOS. It is written primarily in C and x86 assembly language. A revolutionary game, it effectively started the FPS craze that persists to this day. In the game, the player moves throughout a fairly uniform level to reach an elevator that ascends to the next level. Various types of human and non-human adversaries contest the player's progress. The player is armed with four possible weapons: a knife, a pistol, a sub-machinegun and a machinegun. The player's health ranges from 0-100 as does the player's ammunition level (the ammunition is interchangeable with all firearms).

The common enemy in the game is a lowly human guard armed with a pistol containing an infinite amount of ammunition. All of the AAI's in Wolf3D rely on finite-state-machines (FSM) for their behavior. The transitions from one state to another are static. Some randomness is incorporated for AAI targeting accuracy and enemy detection. The randomness influences the speed of transitions in the FSM.

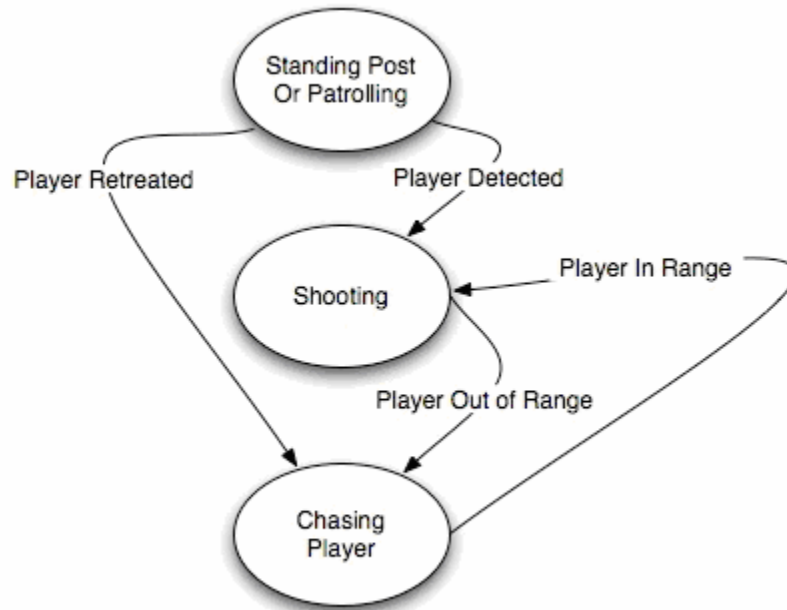


Figure 1. Original state diagram for guard unit

The state diagram for the FSM used to control the guard is shown in Figure 1. From a behavioral standpoint, the guard in the original game could only be described as pitiful, though the supposedly more advanced enemies are hardly any better. AAI's generally charge the player, only stopping at predictable times to fire. There is little notion of self-preservation or tactics. Enhancing the AAI involved three stages: doctrinal enhancements, probability influences and statistically influenced state transitions. Enhancing the doctrine entailed adding new behaviors to the FSM. Probability influences were realized by taking environmental factors into account for determining reaction time. Finally, the FSM transitions were changed to rely on a neural network. The new behaviors added during doctrinal enhancement consisted of evasive shooting behaviors called "strafe-and-shoot" and tactical retreating behaviors. The former behavior entails the AAI constantly moving in random directions that are perpendicular to the player. The AAI pauses briefly to fire at the player. At no time during the maneuver does the AAI close the gap with the player.

Strafe-and-shoot is effectively defensive in nature as it combines evasion with a slightly impaired offensive capability as accuracy is penalized by the constant movement. The tactic of greatest defensive value is the tactical retreat. This retreat involves moving away from the player in a staggered manner. Movement away from the player is interrupted by firing. The AAI is effectively covering its own retreat and hopefully forcing the player to remain at a

distance that prevents a close pursuit. Probability influences were incorporated primarily into the visual system of the AAI. The AAI now detects the player with probability that is influenced by the AAI's position, the lighting level and the acoustics of the room and the AAI's proximity to disturbances such as the sound of gunshots. The modified state diagram is shown in Figure 2. The transitions shown in the figure are influenced by the relative strength of the player compared with the relative strength of the guard.

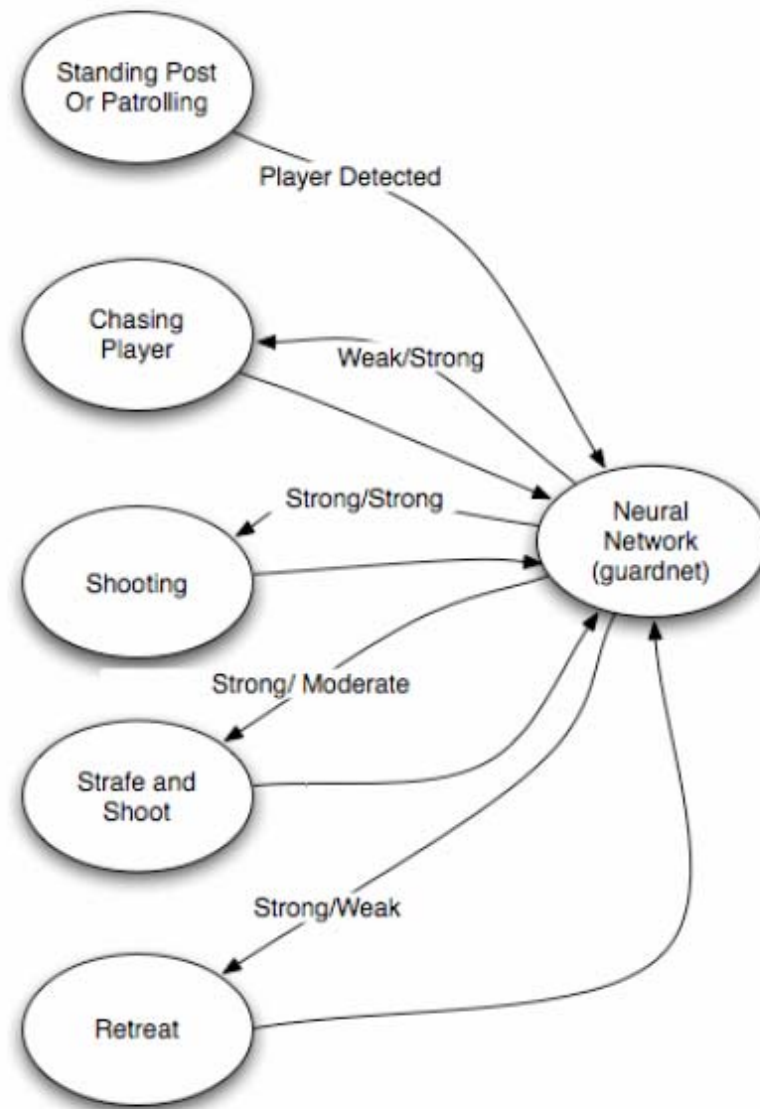


Figure 2. Modified state diagram for guard unit (player / guard)

Lighting and acoustic properties are fixed at compile time with the remaining properties considered at runtime. The original Wolf3D code had probabilities for many of the attributes described above but due to computational constraints of the time they were not incorporated

into the game proper. The probabilities also influence the accuracy of the AAI's observations of the player such as the player's health, weapons choice and ammunition levels. These observations along with the current health level of the AAI are collectively "fed" into the neural network. The neural network is a standalone process that communicates with the Wolf3D executable via the file system. Though not the most efficient inter-process communication mechanisms it is one of the few available in MS-DOS.

The beginning of every behavior contains a function, `agent_think`, that extracts the gamestate information (subject to the probability influences describe above) and writes it to a file. The neural network detects the new file, reads the data and writes a new file containing the behavior that the AAI should next engage in.

The neural network is written in C++. It contains four input nodes, 10 hidden layer nodes and one output node. The network was trained on forty gamestate-decision pairs. As described earlier, each gamestate is represented by a 4-tuple containing the AAI's current health, the current health of the player, the player's current weapon of choice and the player's ammunition level. The decision values range from 0 – 3. 0 corresponds with a decision to retreat. 1 is strafe-and-shoot. 2 is stationary shoot where the AAI comes to a complete halt and makes an aimed shot at the player. Finally, 3 represents advance in which the AAI charges the player's position. The decisions are arranged in a spectrum from defensive to offensive. The training set was intended to be representative of the gamestate space.

The network was a standard feed-forward network and utilized back-propagation for weight-learning. It required about three hours of training to reach an aggregate global error of less than 2×10^{-5} . Output from the network was rounded based on a deterministic model that favored aggression in cases where the AAI had previously performed an offensive action and favored defensive actions when the last behavior was defensive. The neural network did not possess any recurrent properties and the rounding behavior occurred in post-processing. Even with optimizations for the x86 CPU architecture, the neural network was still quite processor intensive and slow relative to the timeslice allocated for each AAI's thinking process. As a result, in some cases the neural network's decision would not be calculated in time, forcing the AAI to continue with its current action until the AAI again invoked the `agent_think` function. The delay, however, in player action and AAI reaction is barely noticeable but still annoyingly visible. Squad behavior was implemented at an elementary level with a 2D positional reservation system. This prevents squad elements from moving to the same 2D position at the same time.

Emergent squad behavior was noted by the playtesters who commented that the various guard elements seemed to be working together or at the least not getting in each other's way. Overall reception of the improvements was good. Most audiences and playtesters found the new version to be "more realistic", "less stupid" and "more challenging." The author, a veteran of Wolf3D, finds the new AAI to be quite formidable and even entry levels involved quite an effort to complete. Even the deterministic rounding model provides for a degree of improvisation. The AAI is unpredictable at times but not a completely random or chaotic fashion.

5. Improvements and Conclusions

Future improvements to the system would include a different inter-process communication mechanism. MS-DOS does come with socket APIs but to interface those

APIs with modern hardware is time consuming and expensive. The better approach would be to integrate the neural network into the Wolf3D executable or perhaps into an external library invoked by Wolf3D. Unfortunately, getting modern C++ code to interface well with archaic MS-DOS code is problematic at best.

The author encountered quite a few limitations with the codebase, particularly the maximum size of a particular module in executable form was severely limited. Wolf3D was initially selected due to the simplicity of environment representation and the 2D nature of the game. Those advantages have to be weighed against the severe limitations imposed upon the developer by the “legacy” nature of the codebase. Implementation issues aside, future improvements to the neural network would be the possibility of learning the player’s habits and tactics, either in real-time or offline based on recordings of humans playing the game.

Recognizing patterns in human behavior and predicting the next behavior would almost certainly entail a recurrent network. Simple pattern observations could be employed where the AAI monitors the player’s health and ammunition levels and based on that tries to determine when the player will disengage from combat and where the player will venture to replenish his/her depleted stores. For more involved squad maneuver combat, recordings of online, multiplayer games (in which all combatants are human players) could be made and winning strategies and tactics analyzed. Obviously, the input vectors for the neural network(s) required to analyze individual and squad behavioral patterns would be enormous, necessitating extensive optimizations. Modern CPU architectures have specific instructions for handling the most common neural network arithmetic operations such as dot-product. Additionally, modern graphics processing units (GPU) have very wide registers that could accommodate long vectors (Hagen, 2005).

Combined with highly optimized versions of the dot-product operation, modern GPUs are orders of magnitude faster at certain operations than even the newest CPUs. The direction of the industry seems to be that of specialized processors and add-on cards. Microsoft’s Xbox-360 has three processors in the same physical package. Combined with its dedicated GPU, the Xbox-360 has four processors each with specialized instructions that are optimized for precisely the type of arithmetic operations that are common to connectionist and probabilistic computational models (Harris, 2005). Recent developments in the PC space include dedicated physics processing units (PPU) incorporated onto PCI expansion cards and the advent of multicore processors similar to those featured in the Xbox-360 (AGEIA, 2005). Game developers will struggle for years to successfully harness this computing power. As a result, the surplus computational resources available can now be used to implement many of the previously computationally intractable connectionist and probabilistic models in real-time. Game developers will have the ability to vastly improve AAI while theoreticians will be able to simulate their models both in real time on a scale previously thought to be unfeasible.

References

- AGEIA Inc. (2005) “A White Paper: Physics, Gameplay and the Physics Processing Unit”.
www.ageia.com
- Busemeyer, Jerome. (2001). Motivational Underpinnings of Utility In Decision Making: Decision Field Theory Analysis of Deprivation and Satiation. Technical Report. Indiana University.
- Carmack, John. (2005). Comments at Quakecon 2005

- Cheong, Mun Hon (2005). Technical Report "Functional Programming and 3D games. University of New South Wales.
- Churchland, P. S., Ramachandran, V., and Sejnowski, T. (1994). *A critique of pure vision. Large-Scale Neuronal Theories of the Brain*. Cambridge, MA: MIT Press.
- Hagen, Trond Runar. (2005). How To Solve Systems of Conservation Laws Numerically Using the Graphics Processor as a High-Performance Computational Engine. Technical Report. SINTEF ICT, Dept of Applied Mathematics.
- Harris, Wil (2005). "How the Xbox 360 affects PC gamers."
http://www.bittech.net/columns/2005/05/13/xbox_360_pc_enthusiasts/1.html
- Hoffman, Achim (2005). "On the Computational Limitations of Neural Network Architectures" Technical Report. University of New South Wales
- Isla, Damian. (2005). Dude, Where's My Warthog? From Pathfinding to General Spatial Competence. AIIDE 2005 Conference.
- Laird, John. (2004). Synthetic Adversaries For Urban Combat Training. *AI Magazine*, 26(3), 82-92.
- Liden, Lars. (2001). The Use Of Artificial Intelligence in the Computer Game Industry. *AI Game Programming Wisdom*.
- Liden, Lars. (2002). Strategic and Tactical Reasoning With Waypoints. *AI Game Programming Wisdom*.
- McBreen, Brendan. (2001). *Squad Size Doesn't Matter*. Marine Corp Gazette.
- McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E. (1955). "A Proposal For the Dartmouth Summer Research Project On Artificial Intelligence".
<http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>
- Orkin, Jeff. (2003). Applying Blackboard Systems to FPS. Presentation UT Austin
- Orkin, Jeff. (2005). Agent Architecture Considerations For Real-Time Planning in Games. AIIDE 2005 Conference.
- Pew, Richard W. (1998). *Modeling Human and Organizational Behavior: Application to Military Simulations*. Washington: National Academy Press.
- Schaeffer, Jonathan. (2001). Gamut of Games. *AI Magazine*, Fall Issue.
- Schieber, M., & Hibbard, L. (1993). How somatopic is the motor cortex hand area? *Science*, 261, 489-492.
- Tozour, Paul. (2005). Spot the Irony. www.ai-blog.net.
- Von Neumann, John. (1944) "Theory of Games and Economic Behavior" Princeton University Press. Princeton, NJ.

Scattered Remarks on Multiple Realizability^s

Hong Yu Wong

Cognitive Science Program, Indiana University Bloomington

*(Adapted with permission from Indiana University Cognitive Science
Program Undergraduate Paper Repository)*

1. Ontological Candor

Since his ontological conversion on the road to supervenience physicalism, John Heil has been proselytizing about ontological candor, and peddling his and Charlie Martin's dual-aspect theory of properties as a somewhat magical cure for "the current malaise in the philosophy of mind" (Heil 1999, 2000). Properties have both dispositional *and* qualitative aspects, and these aspects only come apart in *thought*; in other words, they only come apart in analysis. Heil and Martin will probably balk at my characterization of their theory of properties as "dual-aspect"; they dislike talk in terms of "aspects", for then the qualitative and dispositional "aspects" seem like independent features of properties. Note that on the Heil-Martin theory, properties look just as they would on the pure dispositions account when viewed (or analyzed) from a "dispositional" angle. Heil and Martin tell us no story about why or how properties have both dispositional *and* qualitative aspects, only that properties *have* both dispositional and qualitative aspects. It is a stipulation that comes across as somewhat arbitrary, since the only reason to embrace the dual-aspect theory is its allowing for a physicalistic account of consciousness by fiat.¹ Heil and Martin want the dispositional account, where properties are exhausted by their causal role, and they also want to have properties with intrinsic qualities. They want to diffuse this tension between the dispositional and the qualitative by stipulating that they are but two sides of the same coin. Heil complains of "top-down ontologizing" that his colleagues delight in, yet fails to notice that he is doing exactly the same thing. He is mistaken in thinking that somehow his dual-aspect theory of properties is more adequate than other theories of properties *simpliciter*. The ontology cannot be independently justified; the ends justify the means. But his point about ontological candor is a point well taken.²

Take the multiple realizability intuition for example. There are many questions that one is confronted with once one begins filling in the ontological details. A common construal of mental properties on the functionalist account is as second-order functional properties. On the standard nonreductive physicalist (NRP) account, these second-order functional properties are not epiphenomenal, i.e. they are causally efficacious, but from wherein do they derive their causal powers? If they inherit their causal powers from their first-order realizers why not just reductively identify mental properties with their first-order realizers? The proponent of multiple realizability answers that type identity requires nomological

biconditionals³ which are not available, because mental properties are realized by multiple distinct first-order physical realizers. Each distinct physical realizer is sufficient for the instantiation of the mental property, but none are necessary. Now, depending on one's theory of properties, various difficulties emerge for this account of mental properties. If we take properties to be individuated by their causal contribution (causal powers),⁴ then an apparent contradiction ensues. Imagine a mental property M , say pain (to take a passé example), which is realized distinctly in human, crocodile and octopus neural circuits, N_h , N_c , and N_o respectively. If indeed the first-order physical realizers are distinct, it must be because each physical realizer contributes different causal powers, since we assume properties to be exhausted by their causal roles. So $N_h \blacklozenge N_c \blacklozenge N_o$. But $N_h = M$, $N_c = M$, and $N_o = M$, so by transitivity of "=", $N_h = N_c = N_o$. Apparent contradictions like this evaporate once one adopts a broadly Humean view of causation.

Ontology is the most general architectonic within which we make sense of human activity, inquiry and knowledge. It is how we carve up nature and plumb the fundamental categories. There is a certain relativity to ontological schemes, and different schemes buy you different things. I do not think ontological schemes can be independently justified. Our task is to try out alternative ontologies and find the ontological scheme of best fit. In this task we are constrained by both our pretheoretic intuitions about the world, as well as current scientific knowledge about the world.

In what follows, I will reexamine the centerpiece of nonreductive physicalism—the multiple realizability (MR) argument to irreducibility—and reevaluate its viability *qua* ontology. As with all classic philosophical arguments and examples, MR is given a unique twist by every philosopher who wields it. Some, like Fodor, claim that it is an argument for token physicalism and blocks type reductionism; others, like Kim, have argued that because scientific kinds are individuated by the causal roles they play, and since multiply realized kinds are individuated in the exact same manner, type reduction follows.⁵ To the unsuspecting reader (say a student philosopher), this seems patently contradictory. With a quizzical look she asks, “so multiple realizability leads to both token and type reduction? No?”⁶ Often implicit in discussions of the tenability of NRP and the consequences of MR are the favorite ontologies of individual philosophers (this gets us back to the ontological candor point). Unfortunately (or fortunately), the majority of present day philosophers are closet metaphysicians who maintain a *hush-hush* attitude toward ontological issues. I will attempt to tease out these secret ontologies and show just how crucial these unspoken (and unwritten) ontologies are to each perspective on MR and NRP (and of course to any question in the philosophy of mind). In fact, I like to think that individual ontologies drive individual philosophical enterprises. This should not be a surprise to us: for assumptions (concealed or not) always have consequences.⁸

The structure of the paper will roughly be as follows: I begin by stating the MR contention for philosophy of mind and rehearse why MR poses a problem for type reduction. Philosophers have regarded MR as expressing either an empirical truth or some sort of conceptual necessity; I examine the cogency of these two claims in turn, concentrating on the claim of conceptual necessity. Along the way I also rehearse the intellectual history of MR and reductionist replies and note two varieties of MR: (1) MR across structure types (henceforth, type MR) and (2) MR within a token system (henceforth, token MR). The overriding goal of this paper is to dissect (1) the ontological commitments and (2) theories of causation of nonreductive physicalists and the type reductionists and to evaluate MR within

different ontological frameworks. Disclaimers: In the past decade, a number of philosophers have argued that physicalism is not a well-formed thesis.⁹ I will assume for the sake of discussion that physicalism is a substantive thesis that is concerned with the laws and entities of an ideally completed physics. Nothing essential, however, rides on this assumption.¹⁰ I shall ignore worries about qualia, normativity and agency and will not discuss various alternative frameworks, such as substance dualism. My goal in this paper is to resolve the debate between current reductive and nonreductive physicalists, and so the argument can be read conditionally: *if physicalism is true, what is the most plausible view on the relationship of the mental to the physical?*

2. The Ontology of Multiple Realizability

2.1. MR, Intuitively

Multiple realizability in the philosophy of mind is the contention that a given mental kind (property, state, event)¹¹ can be realized by distinct physical kinds. Putnam was the first to publish this intuition. I once presented this thesis to my biologist girlfriend who promptly rejected it on the grounds that my intuitions were not well founded scientifically. She remained unconvinced after many minutes of hand waving on my part and many fantastic examples ranging from dogs to Martians to robosapiens. I will return to ask whether there is empirical evidence for MR but put this worry aside for the moment. A philosopher's (contentious) apology: philosophy proceeds with a radically different set of normative criteria from science. Intuitions, like MR, drive much philosophizing. What sustains them are sometimes bizarre philosophers' examples (intuition pumps).

Take pain for example (sorry!). It seems plausible that a whole range of animals, from octopuses through humans, are capable of feeling pain, i.e. they are reasonable candidates for pain-bearing types. But consider how vastly different the physiologies and anatomies of these organisms are. Pain, then, must be a multiply realizable kind. Now consider the genius neuroscientist Koch who invents the prosthetic neuron. To prove the efficacy of the Koch neuron, he asks his neurosurgeon colleague Crick to replace one neuron in his primary visual cortex (V1) each day with a Koch prosthetic neuron till his V1 consists entirely of Koch neurons. Koch's trusty graduate students and postdocs perform a daily battery of behavioral tests to see if all functions associated with V1 remain constant. The experiment proves to be a great success. So mental functions strongly correlated with V1 activation are multiply realizable. One can imagine Koch deciding to continue replacing neurons in other parts of his brain with Koch neurons because prosthetic neurons are less susceptible to atrophy due to advanced synthetic materials. He repeats this implantation process a couple times throughout his career, swapping the latest prosthetic neurons for the outdated models. These examples and many other twisted tales of multiply realized minds—the conscious silicon computer HAL, Martians, and other far-flung life forms from the furthest reaches of the universe—sustain the MR intuition.

It is now commonplace to assume that mental kinds are multiply realizable and that MR provided a decisive refutation of type physicalism. The type physicalist thesis is that mental states and processes are type identical to neural states and processes, i.e. there exists a necessary and sufficient physical condition, namely a yet-to-be-discovered neural kind, for the occurrence of every mental kind. On this picture, the relationship between mental and

physical kinds is one-one. If MR is plausible, a given mental kind will have distinct physical realizers, i.e. one-many. Since any of the distinct physical realizers of a mental kind M are sufficient physical conditions for M 's occurrence, and none necessary, there are no necessary *and* sufficient physical conditions for the occurrence of mental kind M , contrary to the assertions of the type physicalist.

2.2. MR, Empirically: Doubtful

There are standardly two ways of understanding the multiple realizability claim. Putnam saw MR as an empirical truth (one that falsified the empirical thesis of mind-brain identity), while later philosophers have regarded MR as expressing a conceptual necessity. I see myself as basically responding to the second construal of MR. It is seldom noted how difficult it is to establish MR as an empirical truth—the first construal—in the first place. There are two worries.¹³ One is the difficulty of individuating psychological kinds in empirical settings. Consider Ron Endicott's papers that have described how mental functions are sustained through neural plasticity, thereby suggesting the multiple realizability of mental states. But this can only be true on the grossest level of grandmother (folk psychological) individuation: "Look, he can still talk even though an iron stake went through his skull!"¹⁴ Furthermore, when we move to consider the difficulties of our projections of exactly the same mental states onto other primates and organisms, we see that the projections which sustain the MR intuition are "untestable" (though there must be some sort of continuity, for otherwise studies of macaque visual systems wouldn't help us understand ours at all). If indeed there is a strong continuity between neural and functional states across species with at least fairly sophisticated level of cognitive function—as evidenced by the reliance of neuroscientists on "psychological measures in mapping the brain and [their doing] so in a comparative fashion" successfully (Bechtel and Mundale 1997, 2000)—then it seems that some sort of identity thesis is far more likely, at least for organic life forms. Perhaps this result pressures us to return to the sort of species-specific-reduction reply that both Lewis and Kim suggest we give to the MR argument to irreducibility.¹⁵ But of course then one can complain that we are exacting an empirical criterion on MR, which is but a philosophical thesis. But remember, we are considering the first interpretation of MR, Putnam's interpretation, where MR is an empirical truth.

2.3. MR, Conceptually: Metaphysical Alternatives

2.3.1. MR under causal realism and the causal powers view of properties

Let us now turn to examine the interpretation of MR as conceptual necessity. If one adopts a sparse ontology—where the only properties admitted into the ontology are those that "cut nature at its joints"—and individuates properties on the basis of their causal powers,¹⁶ then MR amounts to no more than a second-order manner of picking out first-order entities. This is indeed the functionalist construal of mental properties.¹⁷ One must then ask what the ontological status of these second-order properties is.¹⁸ Let us now turn to examine the possible cases.

Epiphenomenalism.

If second-order properties are epiphenomenal, as Block (1990, 1995) suggests, then mental properties contribute nothing *qua* causal and since properties are exhausted by their causal powers, there are no *effective* mental properties and the question of reducibility or irreducibility dissolves. All that exists are physical properties, since only these are causally potent. On this picture, mental predicates are linguistic locutions having positive epistemic status that play useful roles in the communication and cognition (folk psychology) of human cognizers, but no more.

Token identity.

If second-order properties pick out psychological properties, and if they inherit the causal powers of their token physical realizers, then their causal powers are nothing over and above the causal powers of their token physical realizers. These second-order properties are then identical to the first-order realizers since we individuate properties based on their causal contribution. Furthermore, if the first-order physical realizers are distinct, as is standardly assumed, then they will be in each case distinct properties. There is a tension in holding the token identity of the mental property in question with each of its distinct physical realizers: imagine a mental property M, which is realized distinctly in human, crocodile and octopus neural circuits, N_h , N_c , and N_o respectively. If indeed the first-order physical realizers are distinct, it must be because each physical realizer contributes different causal powers, since we assume properties to be exhausted by their causal roles. So $N_h \blacklozenge N_c \blacklozenge N_o$. But $N_h = M$, $N_c = M$, and $N_o = M$, so by transitivity of “=”, $N_h = N_c = N_o$. To escape contradiction ($N_h \blacklozenge N_c \blacklozenge N_o$ and $N_h = N_c = N_o$), the only consistent way of interpreting the second-order mental property is as a predicate which does not figure in ontology, i.e. as an epistemic-linguistic locution that picks out first-order properties. But we can have as many of these linguistic entities as we want; jade is an excellent example.¹⁹

(Note the differences between Davidson-style and Kim-style events and difficulties with Davidsonian events. It is important to realize that one’s ontology of events can impact one’s understanding of token identity. In particular, differences between Davidson and Kim events are central to the issue of reducibility or irreducibility when one espouses a token identity relation. Kim understands events as property exemplifications at specific times, whilst Davidson understands events as concrete particulars that can fall under event-kinds. On Davidson’s account of events, token identity holds when an event falls under both a mental event-kind and a physical event-kind. On this account of token identity, the relationship between mental and physical properties is hopelessly obscure. Why so? Consider an object which has size and has shape. What is the relationship of size and shape? Is it one of identity? We know it is not. But in general, the relationship between two properties of an object may not be deduced simply from the fact that an object has both those properties. Analogously, even though a concrete event may fall under both mental and physical event-kinds, the relation of the mental and physical aspects of the event is unclear. What I think Davidson’s account amounts to is agnosticism with regard to the mind-body relation. Throughout this paper, I will be assuming Kim’s account of events. See pp. 58-62, Kim (1996) for a discussion of these issues. See also LePore and Loewer (1987).)

Running the above line on the ontological status of second-order mental properties requires a story about properties of ordinary objects, for otherwise one would be vulnerable

to a *reductio* claiming that the exact same worries about mental causation generalizes to anything which is not a fundamental physical property.²⁰ Here's a quick story: We are to understand the properties of complex physical states as being the composition of all²¹ their microphysical properties. In everyday experience, ordinary objects appear to have unique causal powers—e.g. a square peg with 1 inch sides cannot fit in a round hole with a 1 inch diameter—because the composition of microphysical properties results in apparently unique properties to human observers. However their properties *just are* the composition of the microphysical realizers (i.e. these are what Armstrong calls structural properties) and are nothing over and above the properties of their microphysical constituents. The apparent novelty of causal contribution is yet another epistemic artifact.

Nonstandard views of realization.

We now turn to consider an alternative view of realization which purports to allow for MR and irreducibility. Sydney Shoemaker sets out his most recent view of realization in his (1999), wherein properties are individuated by their causal features.²² This is cashed out in terms of what he calls “forward-looking and backward-looking causal features” which allows for more fine-grained property individuation than the token identity view. The forward-looking causal features of a property P are the causal powers that are had by P (“... what contribution their instantiation can make to causing various effects”), while the backward-looking causal features are the set of causal features of states of affairs that might cause P's instantiation. So on Shoemaker's picture, if property P realizes property Q, then the “forward-looking causal features of Q are a subset of the forward-looking causal features of P, and the backward-looking causal features of P are a subset of the backward-looking causal features of Q” (see diagrams 1 and 2). Notice that this allows that “it is sometimes the realized property, not its realizer, that causes a certain effect; this will be so when the causal features involved in the causal transaction belong to the subset which the realized property shares with all its realizers.”

So on Shoemaker's picture, a property *M* is multiply realizable and hence irreducible since *M*'s forward-looking and backward-looking are not identical with any of its distinct neural realizers, *N*₁, *N*₂, ..., *N*_n. Notice, however, that since properties are to be individuated on the basis of their forward-looking and backward-looking causal features, nothing prevents us from identifying property *M* in the diagrams below with causal features *P*₁, *P*₂, and *P*₃, which we identify as property *N*'. For example, we can construe property *N*₁ as *N*' plus *P*₄, and property *N*₂ as *N*' plus *P*₅. At this juncture, it seems that Shoemaker has two options: (1) dismiss *N*' as an empirically nonexistent property, or (2) identify *M* with *N*', and understand *M* as a second-order linguistic locution that picks out its distinct first-order realizers {*N*₁, *N*₂, ..., *N*_n}. Because Shoemaker chooses to individuate properties on the basis of their causal features,²³ he may not reply to the reduction of *M* to *N*' = {*P*₁, *P*₂, *P*₃} by claiming that the identification fails because *N*' does not exist empirically since *N*' just is the causal features *P*₁, *P*₂, and *P*₃.²⁴

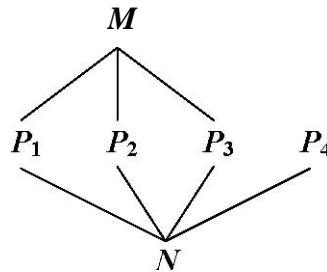


Figure 1. A property N realizing M in Shoemaker's picture.

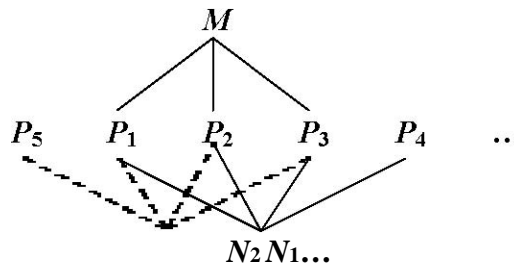


Figure 2. Multiple realizability in Shoemaker's picture.

In all the cases above, the multiple realizability amounts to no more than a second-order linguistic locution that picks out or names the set of properties $\{P_1, P_2, \dots, P_n\}$.²⁵ This is certainly insufficient to argue for irreducibility. The question of reducibility does not even arise in these contexts, since the supposed mental properties fail to qualify for propertyhood. So the only way that there could be a relationship of irreducibility between mental properties and their physical realizers is if one conceals a property dualism to begin with; the multiple realizability argument to irreducibility is thus circular. Functionalism (the NRP version) then fails to be established. Functional kinds can however be identified with their physical realizers as in Lewis-Armstrong causal functionalism.²⁶ This is just as in the carburetor case. It is not the property of being a carburetor that does the causal work. It is the internal workings of the X, Y and Z parts of the carburetor and their interactions that allow it to play the causal role that is required. One can call those things "carburetors" if one likes.

2.4. MR under a deflationary account of causation

Turn now to consider multiple realizability and nonreductive physicalism under a broadly Humean, deflationary perspective toward causation. Hume's view is that there is nothing

intrinsically intelligible about causality, and that there are no necessary connections between a cause and its effect. “The contrary of every matter of fact is still possible; because it can never imply a contradiction” (Hume 1748, p. 25). Our innate intuitions concerning intrinsic necessary connections existing between causes and effects are due to our recognizing regularities in our experience where the members of a class of events and objects are constantly conjoined with members of another class of events and objects and our projecting an *intrinsicness* into the causal relation. Hume further suggested that the truthmakers of purported causal relations are the associated counterfactual conditional statements. He wrote: “We may define a cause to be *an object, followed by another, and where all the objects, similar to the first, are followed by objects similar to the second*. Or, in other words, *if the first object had not been, the second never had existed*” (Hume 1748, P. 76). Hume did not go on to develop a counterfactual analysis of causation, and it was not until the 1970s that counterfactual treatments—largely owing to the seminal work of David Lewis²⁷—became immensely popular.²⁸

It turns out that under a Lewisian counterfactual analysis of causation, the causal efficacy of higher-level properties is not an issue at all (and we should expect overdetermination). Proponents of nonreductive physicalism standardly appeal to counterfactual accounts of causation where every cause/effect pair instantiates a law. NRP defenders typically also espouse a rich (almost profligate) ontology, where any predicate that is subsumed under a law (regardless of whether the law is a strict law or a *ceteris paribus* one) is a causally potent property.²⁹ Fodor’s line of thought on the predicate property distinction is that any predicate that is subsumable under law exists and is a causally potent property. He charges Kim with making an unnecessary constraint on properties: that they must be subsumable under physical law, which is why higher-order properties on Kim’s count turn out to be causally impotent. Fodor himself is agnostic on metaphysical issues³⁰ but has acknowledged that if one adopts a Humean attitude toward causation, many of the worries evaporate.³¹ The counterfactual defense of NRP has been undertaken by numerous philosophers who espouse some sort of Humean supervenience of laws on the set of local, intrinsic facts, including Block (1990), Horgan (1989, 1993, and forthcoming), LePore and Loewer (1987, 1989), and Loewer (forthcoming). Among these, Horgan’s account is the most carefully worked out one, and I will sketch his defense against arguments for the causal exclusion of higher-order mental properties below. Horgan’s account, like all other defenses of NRP, relies heavily on the notion of explanatory relevance. Horgan begins in his (1993) by giving a careful account of causal explanation. Consider the case where we want to give an explanation of the causal transaction between a phenomenon *c* and a phenomenon *e*, where *c* is instantiating a property of type *C* and *e* is instantiating a property of type *E*. For properties *C* and *E* to be genuinely explanatorily relevant to the causal transaction between phenomena *c* and *e*, Horgan, following Woodward (1979), requires not only that *c* caused *e* and that *c* and *e* are subsumable under a counterfactual, law-like generalization (this satisfies the Fodor propertyhood requirement), but importantly also that *C* and *E* must fit into a suitably rich pattern of counterfactual relations among properties. Horgan calls this the counterfactual pattern conception (CP) to explanatory relevance.

Under the CP conception, a single phenomenon, e.g. that *c* caused *e*, can be subject to a variety of equally valid explanations involving different theoretical levels of counterfactual, law-like generalizations. Typically, the level of theoretical explanation is discourse sensitive and determined by the relevant level of detail. Ignoring worries about the *ceteris paribus*

nature of psychological laws,³² each level of counterfactual generalizations is equally explanatorily valid for the context of discourse in question. Consider a certain counterfactual generalization about neurophysiological states that relates a state N_1 to a state N_2 . Even if multiple realizability doesn't hold, explanation in terms of the conjunction of huge numbers of quantum mechanical states that make up complex molecular structures that in turn make up the neurophysiological states seems infeasible. Certainly, just churning the numbers out (most likely with the help of significant computational aids) would suffice for an instrumentalist (or behaviorist) analysis, but substituting the vast quantum-state conjunctions for a neurophysiological predicate in a covering law would scarcely be "explanatorily adequate". The law wouldn't make any sense. So at each grain of analysis, there is an explanatorily appropriate level of counterfactual, law-like generalizations (which, as you might remember, are embedded in some rich network of counterfactual dependencies). Remember also that any predicate that figures in a counterfactual, law-like generalization names a causally efficacious property, so Horgan's Humean account gives us a rich, layered cake of properties.

3. Humean Worries: Irrealism, Inter-Level Constraints and Ontic Primacy

An immediate objection against Horgan's account is that explanation is an anthropocentric epistemological notion, and fails to speak to the ontological worries at hand. Horgan is aware of this objection,³³ but fails to grasp the full extent of the ontological problem. In fact, he makes a fatal error: he implicitly conflates explanatory relevance with causal relevance. After his explication of the CP conception, Horgan rushes to defend CP against what Jaegwon Kim calls *explanatory irrealism*, which Kim describes as "the view that the relation of being an explanans for, as it relates C and E within our epistemic corpus, is not, and need not be, 'grounded' in any objective relation between events c and e ".³⁴ Defending the Humean account against Kim's explanatory irrealism charge is easy, as the irrealism charge is far too strong. Horgan's response to Kim's charge is appropriate: indeed an advocate of the CP conception can point to objective grounding for the explanans/explanandum relation, for on the NRP count, causal closure of physics implies that for every event there is a complete causal explanation in terms of physical properties. The nonreductive physicalist then tells a story where the complete physical causal story doesn't trump counterfactual, law-like causal generalizations at other levels of explanation, but instead provides the ontological grounding for all (remember that the nonreductive physicalist espouses a position of token identity with respect to the mental-physical relation³⁵) other levels of causal explanation. However, Horgan still needs to respond to the charges that higher level explanatory relations are neither objective nor mind-independent.³⁷ Remember that the notion of explanation is "making sense of"—it is a cognitive notion and hence is relative to a class of cognizers—and that the relevance and efficacy of explanation is discourse and grain dependent. He has given us no reason to think the explanatory relevance implies causal relevance. His assertion that predicates figuring in explanations automatically acquire propertyhood is suspect; the tail is wagging the dog. What the arguments of Horgan (and Putnam³⁸ and Fodor³⁹ before him) effectively establish, is only that certain ways of characterizing reality are essential in helping human beings to make sense of the world, and these high-level vocabularies are, for all practical purposes, ineliminable.

There are further worries about the NRP metaphysical framework. On the CP conception, a single phenomenon, e.g. that c caused e , can be subject to a variety of equally valid explanations involving different theoretical levels of counterfactual, lawlike generalizations. Furthermore, these nomic counterfactual generalizations are embedded in a larger complex network of nomic counterfactual generalizations that is the “web of science”, to use a Quinean term. We might ask what the interlevel constraints on this network of counterfactual levels are. We might think that for nonreductive physicalists the levels of nomic counterfactual generalizations are equally valid explanatorily, and hence properties at different levels of the layered ontology that NRP espouses have equal status. There is no question of ontic primacy of the physical; physical and mental properties are ontologically equally entities. To remedy this situation, NRP adds the assumption of causal closure. We might also ask what lines up the counterfactuals at different levels of explanation perfectly, such that physical, chemical, biological and mental explanations are equally valid and compatible. The story that Horgan provides us with is that everything else synchronically supervenes on the physical. But notice that unless NRP is espousing some sort of Leibnizian preestablished harmony (which nonreductive physicalists certainly don’t want) and unless Horgan is conflating explanatory relevance with causal relevance (or epistemological relevance with ontological relevance), the process of gathering nomic counterfactual generalizations (also known as “doing science”) is a messy diachronic process and it is by no means clear that counterfactual generalizations across levels will *always* be mutually consistent. There are two complaints that I have: (1) NRP seems to carve up scientific levels of explanation too neatly and discretely, but in reality the levels of explanation are continuous, e.g. chemical physics and physical chemistry, or biochemistry and molecular biology. (2) In fact, there is also good reason to doubt the autonomy of the upper ontological strata on the NRP picture because there are interlevel constraints that involve the day-to-day process of doing science. Ask any good scientist and he or she will tell you that each level of scientific explanation must be consistent with its adjacent levels. Thus, there must be some interlevel traffic, which is what allows knowledge acquired in physics to filter up to molecular biology. A good example of this is how advancements in crystallography allowed for the elucidation of DNA structure. This filtering of knowledge is not transitive *qua* cognizers (an evolutionary biologist will *not* care about advancements in atomic theory), but since each level *must* be consistent with its adjacent levels (local consistency), knowledge acquired filters up the levels of explanation (there is a pressure toward global consistency).⁴⁰

As I noted earlier, under the CP conception, each level of explanation is *equally valid*, hence predicates that figure at different levels of explanation are *ontologically equal*. The ontological primacy of the physical that is so much a part of the physicalist picture does not figure in the CP conception as such and needs to be further stipulated; I find this somewhat unnatural. I would imagine that for physicalists (the “P” of the NRP) physical properties should have pride of place in their ontology and this ontological primacy should be a natural part of their metaphysical framework. Typically, the ontic primacy of the physical is hidden either in the realization relation, or some further stipulation like Lewis’ notion of naturalness. For example, Loewer and LePore in their (1987) require that an event e ’s being X must explain e ’s being Y if e ’s being X is to realize e ’s being Y . This explanatory requirement is meant to be stronger than plain physical necessity. [I find the realization relation somewhat suspect. Realization talk in the philosophy of mind was established in the 1960s by a series of early Putnam papers on the strength of computational analogies (abstract mathematical machines are realized by concrete physical devices). It has been noted more than once that

the notion of realization has largely been taken for granted, and little or no work has been done to explicate the notion with regard to the more traditional options on the mind-body problem.⁴¹ Furthermore, it is not even clear that we understand the realization relation in the case of computers.⁴²]

4. Type and token MR

In his paper on “Multiple Realizability, Multiple Reference and the Reduction of Mind”, Horgan charges that David Lewis’ reductive causal functionalist account—which my account has broad sympathy with—fails to accommodate what Horgan terms strong multiple realizability. Horgan’s objection rests on a technicality based on Lewis’ use of the notion of a creature-kind in his explication of mental concepts as nonrigid designators. It is standard in the philosophy of mind literature to differentiate between two varieties of MR: type and token (my terms). Type MR is the idea that a mental property is realized by distinct structure types; token MR, which has been called radical MR (Polger) or strong MR (Horgan) in the literature, is the idea that a mental property might be realized by distinct physical tokens in a token system across times. As Horgan points out, Lewis’ treatment of mental concepts as nonrigid designators having multiple referents flounders because he implicitly restricts the reference of mental terms to a creature-kind (remember the Lewis-Kim species-specific reduction reply). This is not a problem for Lewis’ ontology, for Lewis can drop the restriction of reference being restricted to creature-kinds and further narrow the reference to token systems at specific times. One might complain that in making this move psychological predicates lose their generality and this poses problems for concocting psychological theories. But remember that we are concerned with the ontological status of mental predicates, not their epistemological status or utility in human discourse.

5. What Remains

The metaphysical groundwork and argumentative details in this paper are woefully incomplete at best, but I hope I have shown the importance of ontological candor in theorizing about the mind and provided the skeleton of arguments that should raise grave doubts about the multiple realizability argument to irreducibility qua ontology. I have no quarrel with the need for higher level vocabularies (with positive epistemic status) that pick out lower level realizers in our ongoing pursuit for knowledge about the world. In fact, if Putnam and Fodor’s antireductionist arguments have shown anything, they have shown that such higher level vocabularies, such as mental discourse, are practically ineliminable. But if my arguments are sound, then these predicates fail to pick out fundamental properties. Perhaps this indicates that the only way to be a mental realist is to renounce our physicalist commitments and embrace some sort of robust dualism, but that remains to be seen.⁴³

Notes

\$. Submitted for the CUNY 5th Annual Graduate Student Philosophy Conference, 2001.

1. I am perhaps being unfair to Heil and Martin here. Unfortunately, the only example Heil applies the dual-

aspect theory to in his published corpus is that of consciousness. Perhaps Heil can say in defense that because scientific kinds are generally causal kinds, science has tended to characterize physical properties functionally, and hence we have no good examples of qualitative aspects of physical properties.

2. On ontological candor, see Heil and Martin (1998), Robb and Heil (1998) and Stewart (1997).

3. It is unclear that nomological biconditionals are necessary for type reduction unless one accepts the classical Nagel model of reduction. Jaegwon Kim has argued that the Nagel model is neither necessary nor sufficient for reduction (Kim, 1998, chapter 4). In particular, the existence of Nagelian bridge laws connecting a reducing and a reduced theory is consistent with the doctrine of the British emergentists, who allowed for nonexplanatory, brute bridge laws which were to be accepted with “natural piety”. If only nonexplanatory bridge laws link a reduced and a reducing theory, then the reduced theory has not really been reduced, since the reduced theory cannot be reductively explained in terms of the reducing theory.

4. See Shoemaker (1980).

5. This is the gist of the functional model of reduction that Kim currently espouses. Jaegwon Kim began as one of the foremost proponents of supervenience physicalism, the doctrine that mental supervenes on the physical but is irreducible to the physical. He has since moved away from nonreductive physicalism toward a robust type-reductionism based on a functional model of reduction (Kim 1998, 1999). See Horgan (1996) for a review of Kim’s positions on the mind-body problem through Kim (1993) and essays in the section “Mental Causation, Reduction and Supervenience” (pp. 83-208) in *Philosophical Perspectives*, Vol. 11 (1997) for reviews of Kim’s particular brand of functionalism.

6. The resolution of this issue hinges on one’s exact ontology of events. I will return to consider differences between Davidsonian and Kimian events later in this paper, and the relation between their ontology of events and their solution to the mind-body problem.

7. This reminds me of Jerry Fodor’s quip in his (1985, p. 76): “It rained for weeks and we were all so tired of ontology, but there didn’t seem to be much else to do.”

8. Compare John Heil’s comments in his (2000): “I am convinced, however, that issues now at the forefront of the philosophy of mind are fundamentally metaphysical in character. (This is scarcely surprising. The philosophy of mind is, after all, a kind of applied metaphysics [emphasis not in original]. It is crazy to think that philosophers of mind can ignore or remain neutral on questions of ontology.)”

9. See Crane and Mellor (1990) for arguments to this conclusion.

10. I will also ignore empirical worries concerning the discontinuities between various varieties of MR. Polger (ms.) differentiates between MR in cases where the realizers are somewhat similar (say the realization of the visual system in higher primates) and where the realizers are radically distinct (say Martians and green cheese and silicon androids all realizing pain).

11. [The ontology of these metaphysical entities is as yet an unsettled issue (Stewart, 1997).]

13. See Zangwill (1992) for further discussion.

14. This is, of course, a caricature—but one that contains a kernel of truth.

15. See pp. 233-236, Kim (1996), Kim’s comments on the structure-restricted correlation thesis in his (1992), and Lewis (1969).

16. Shoemaker gives a robust account of this in his (1980), while Kim has often assumed a similar individuation criteria that he has christened “Alexander’s dictum”: a property is real if and only if it has causal powers. Note that Kim’s individuation criterion is weaker than Shoemaker’s.

17. The relation between functionalism and multiple realizability is fuzzy. One can easily imagine examples of functionally defined kinds which are not multiply realizable. Consider a functional characterization of any elementary particle, in term of the causal-functional role in high energy physics interactions. Even though the elementary particle in question can be functionally defined, it is not multiply realizable.

18. Note that functional properties are second-order properties.

19. See Kim (1992), pp. 11-19.

20. See Kim (1997), Noordhof (1999), Kim (1999), Kim (1998), Block’s (ms.), Loewer (2001), and Menzies (2001).

21. [Idea of “composition” needs to be stated more carefully. Ontological issues impinge: structural universals? tropes? or a linguistic view of compositionality (the whole is a function of its parts)?]

22. It is not immediately clear to me what causal features are. Are these causal powers? Are they finer grained? Which individuation criteria are we to contrast the causal features view against?

23. See Shoemaker 1980.

24. I not here consider another view of realization, somewhat similar to Shoemaker’s where instead of M’s causal features being a subset of M’s realizers’, M’s realizers’ causal features are a subset of M’s. Multiple

realizability on this picture would look like this: assume M is realized by two distinct neural properties N1 and N2. $M = \{P1, P2, P3, P4, P5, P6\}$, $N1 = \{P1, P2, P3, P4\}$, and $N2 = \{P1, P2, P3, P5\}$. Now, M is not reducible to either of its realizers, but someone who espouses such a view would have to explain how P6 emerges mysteriously. My treatment of Shoemaker's picture in this section is heavily indebted to Heil's treatment in his (1999).

25. In his (1999), Kim arrives at a similar conclusion. He suggests that we "give up [the higher-order mental property] E as a genuine property, only recognizing the expression "E" or the concept of E. As it turns out, many different properties are picked out by E, depending on the circumstances One could argue that by forming "second-order" functional expressions by existentially quantifying over "first-order" properties, we cannot be generating new properties, only new ways of indifferently picking out first-order properties in terms of certain causal specifications that are of interest to us (p. 17)." See also Heil (1999) & Antony (1999), pp.1-9

26. I am not claiming here that Lewis or Armstrong's metaphysical views and their philosophies of mind are indeed consistent. This is not true, especially in the case of Lewis. Lewis is a Humean about causation, and as we shall soon see, there is a great tension between his causal functionalism and his Humean counterfactual treatment of causation. Terry Horgan recently noted this point in his forthcoming essay "Multiple Reference, Multiple Realization, and the Reduction of Mind".

27. See Lewis (1973).

28. Hume actually was confusing two disparate deflationary accounts of causation, the regularity theory and the counterfactual analysis.

29. I.e. every predicate that figures in a covering law explanation corresponds to a property. There is a further implicit assumption that any explanatorily relevant property is also causally relevant.

Epiphenomenal properties are explanatorily irrelevant and hence do not figure in the nonreductive physicalist's account.

30. Fodor is apathetic about physicalism almost to the point of agnosticism. He writes in his "Making Mind Matter More": "I'm not really convinced that it matters very much whether the mental is physical; still less that it matters very much whether we can prove that it is. Whereas, if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my saying ... if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world." (Fodor, 1989, p. 77)

31. Fodor, in conversation.

32. The claim that the lower-level covering laws of physics are "strict" and "exceptionless" is quite preposterous. Quite the contrary, physical laws, like Newton's laws and Hooke's law hold only *ceteris paribus* as well. Newton's laws for example only hold at velocities radically smaller than the speed of light. Having said that, we expect the fundamental field equations (or string theoretic equations for that matter) of completed physics to be exceptionless.

33. Horgan writes: "This kind of context/purpose relativity is entirely compatible, as far as I can see, with the contention that facts about explanatory relations are objective and mind-independent. Hence the CP conception does not embody any commitment to what Jaegwon Kim calls explanatory irrealism." (Horgan 1993, p. 300)

34. Kim (1988), pp. 226-227.

35. In general, the nonreductive physicalists espouse token identity as the relation between physical properties and any other properties from higher levels in the layered NRP picture of the world. There are also other varieties of nonreductive physicalism where the relationship between higher-order properties and physical properties is not one of token identity, but global supervenience. Such accounts are often motivated by concerns regarding wide content or externalism. Note also that the nonreductive physicalist espousing such a global supervenience account needs to give an account of what it means for a set of properties to globally supervene on another set of properties beyond just stating definitions characterizing some sort of property-dependent modal covariation. See McLaughlin (1995) for an exhaustive treatment of supervenience. See Horgan (1993) and Kim (1998) for arguments to the conclusion that supervenience physicalism fails to characterize a distinctive theory of mind, and that consumers of supervenience need to further explain the property-dependent modal covariation.

36. At the moment I simply want to note that there is a tension in the token identity view. Giving physical properties the pride of place in our ontology, whilst maintaining that properties at other levels in the layered ontology have separate ontological status, leads to numerous problems. These I will discuss soon enough.

37. He needs to confer propertyhood to properties at higher levels in the layered ontology, while at the same time grounding them in the physical realm.

38. See Putnam's "Philosophy and our Mental Life" (1973), in *Mind, Language and Reality* (1975), pp. 291-

303.

39 See Fodor (1974).

40 Consider the possibility of interlevel counterfactual generalizations.

41 See Kim (1998) pp. 7-9, Shoemaker (1999), and Block's discussion of the realization relation in his (1995).

42 [Brian Cantwell Smith has repeatedly emphasized this point to me in conversation. According to Smith, the mind-body problem for computers is not trivial, and neither is the concrete/abstract distinction; one does not get notions like realization and the concrete/abstract distinction for free from the realm of computing. Philosophical analysis has yet to be performed on such basic notions in computing such as implementation and realization. See Sloman (1998) and Smith (1996).]

References

- Louise Antony (1999), "Multiple Realizability, Projectibility, and the Reality of Mental Properties", *Philosophical Topics*, 26, pp. 1-24.
- William Bechtel and Jennifer Mundale (1997), "Multiple Realizability Revisited", *Proceedings of the Australian Cognitive Science Society*, URL = <http://artsci.wustl.edu/~bill/multiple.htm>
- (1999), "Multiple realizability revisited: Linking cognitive and neural states", *Philosophy of Science*, 66, pp. 175-207.
- John Bickle (1998), "Multiple Realizability", *The Stanford Encyclopedia of Philosophy* (Spring 2001 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/entries/multiple-realizability/>
- Ned Block (1990), "Can the Mind Change the World?", in Boolos (ed.), *Mind and Method*, Cambridge: Cambridge University Press, pp. zz-zz.
- (1995), "What is Functionalism?", in Guttenplan (ed.), *A Companion to the Philosophy of Mind*, pp. xxx-xxx.
- (ms.), "Do Causal Powers Drain Away?", manuscript.
- Tim Crane and D. H. Mellor (1990), "There is No Question of Physicalism", *Mind*, 99, pp. 185-206.
- Jerry Fodor (1974), "Special Sciences, or The Disunity of Science as a Working Hypothesis", *Synthese*, 28, pp. 97-115.
- (1985), "Fodor's Guide to Mental Representation: The Intelligent Auntie's Vade-Mecum", *Mind*, 94:373, pp. 76-100.
- (1989), "Making Mind Matter More", *Philosophical Topics*, 17, pp. 59-79..
- John Heil (1999), "Multiple Realizability", *American Philosophical Quarterly*, 36:3, pp. 189-208.
- (2000), "Metaphysics of Mind", *A Field Guide to the Philosophy of Mind*, Nani and Marraffa (eds.), URL = <http://www.uniroma3.it/kant/field/mm.htm>
- John Heil and C. B. Martin (1999), "The Ontological Turn", *Midwest Studies in Philosophy*, 23, pp. 34-60.
- Terry Horgan (1993), "Nonreductive Materialism and the Explanatory Autonomy of Psychology", in Wagner and Warner (eds.), *Naturalism: A Critical Appraisal*.
- (1996), "Kim on the Mind-Body Problem", *British Journal for the Philosophy of Science*, 47, pp. 579-607.
- (2001), "Multiple Reference, Multiple Realization, and the Reduction of Mind", forthcoming in Sibel and Preyer (eds.), *Reality and Humean Supervenience: Essays on the Philosophy of David Lewis*.
- David Hume (1748), *Enquiry Concerning Human Understanding*, Selby-Bigge (ed.), Nidditch

- (rev.), 3rd edn, Oxford: Oxford University Press, 1986.
- Jaegwon Kim (1988), "Explanatory realism, Causal Realism, and Explanatory Exclusion", *Midwest Studies in Philosophy*, 12, pp. 225-240.
- (1992), "Multiple Realization and the Metaphysics of Reduction", *Philosophy and Phenomenological Research*, 52, pp. 1-26.
- (1993), *Supervenience and Mind*, Cambridge: Cambridge University Press.
- (1996), *Philosophy of Mind*, Boulder: Westview Press.
- (1997), "Does the Problem of Mental Causation Generalize?", *Proceedings of the Aristotelian Society*, 97, pp. 281-297.
- (1998), *Mind in a Physical World*, Cambridge: MIT Press.
- (1999), "Making Sense of Emergence", *Philosophical Studies*, 95, pp. 3-36.
- David Lewis (1969), "Review of Putnam", reprinted in Block (ed.), *Readings in the Philosophy of Psychology*, vol 1, pp. 232-233, 1980.
- (1973), "Causation", *Journal of Philosophy*, 70, pp. 556-567.
- (1995), "Lewis, David: Reduction of Mind", in Guttenplan (ed.), *A Companion to the Philosophy of Mind*, pp. 412-431.
- Ernst LePore and Barry Loewer (1987), "Mind Matters", *Journal of Philosophy*, 84, pp. 630-642.
- (1989), "More on Making Mind Matter", *Philosophical Topics*, 17, pp.175-191.
- Barry Loewer (2001), "Review of Kim's *Mind in a Physical World*", manuscript, forthcoming in *Philosophy and Phenomenological Research*.
- Brian P. McLaughlin (1995), "Varieties of Supervenience", in *Supervenience: New Essays*, Savellos and Yalçın (eds.), pp. 16-59, Cambridge: Cambridge University Press.
- Peter Menzies (2001), "The Causal Efficacy of Mental States", in Monnoyer (ed.), *The Structure of the World: the Renewal of Metaphysics in the Australian School*, Vrin Publishers.
- Thomas Polger (2000), "Putnam's Intuition and Multiple Realizability", manuscript.
- Hilary Putnam (1973), "Philosophy and our Mental Life", in *Mind, Language and Reality* (1975), pp. 291-303, Cambridge: Cambridge University Press.
- David Robb and John Heil (1998), "Ontology and Mental Properties", manuscript.
- Sydney Shoemaker (1980), "Causality and Properties", in van Inwagen (ed.), *Time and Cause: Essays Presented to Richard Taylor*, pp.109-136, Reidel: Dordrecht.
- (1999), "Realization and Mental Causation" in Elevitch (ed.), *Proceedings of the 20th World Congress of Philosophy*, Vol. IX: Philosophy of Mind, Bowling Green: Philosophy Documentation Center, Bowling Green State University.
- Aaron Sloman (1998), "Supervenience and Implementation: Virtual and Physical Machines", Virtual and Physical Machines Technical Report, School of Computer Science, University of Birmingham.
- Brian Cantwell Smith (1996), *On the Origin of Objects*, Cambridge: MIT Press.
- Helen Stewart (1997), *The Ontology Of Mind: Events, Processes, and States*, Clarendon Oxford Press: New York.
- James Tomberlin (ed.) (1997), *Philosophical Perspectives*, Vol. 11: Mind, Causation, and World.
- Peter van Inwagen (ed.) (1980), *Time and Cause: Essays Presented to Richard Taylor*, Reidel: Dordrecht.
- James Woodward, (1979), "Scientific Explanation", *British Journal for the Philosophy of Science*, 30:41-67, pp.54-55.
- Nick Zangwill (1992), "Variable Realization: Not Proven", *The Philosophical Quarterly*, 42:167, pp. 214-219.

Indiana Undergraduate Journal of Cognitive Science

2006 – 2007 Editorial Board

Executive Editor

Michael T. Amlung, *Indiana University Bloomington*

Associate Editor

Elton Joe, *Hampshire College & Indiana University Bloomington*

Student Reviewers

Sarah Coleman, *Indiana University Bloomington*

Jordan DeLong, *Indiana University Bloomington*

Melissa Troyer, *Indiana University Bloomington*

Faculty Sponsors

Dr. Ruth Eberle, Assistant Professor of Cognitive Science & Informatics,
Indiana University Bloomington

Dr. Robert Goldstone, Chancellor's Professor and Director of Cognitive Science
Program, *Indiana University Bloomington*

Web Design / Programming Support

Fang Fang, *Indiana University Bloomington*

Indiana Undergraduate Journal of Cognitive Science

<http://www.cogs.indiana.edu/iacs/journal.html>

Author / Submission Instructions

I. General Information

The Indiana Undergraduate Journal of Cognitive Science invites submissions of original writing by undergraduate students. Submissions may come from any area within Cognitive Science including, but not limited to: artificial intelligence, anthropology, biology, computer science, linguistics, philosophy, psychology and neuroscience.

II. Submission / Paper Format

Articles are accepted on a continuous basis and will be considered for publication upon submission. Articles should be sent directly to the editorial board as an attachment in Microsoft Word or Adobe PDF format. Submissions should be edited for grammar and style before submission. There is no limit on article length. Submissions should include a Title Page that includes the following information: Article Title, Author Name, Major, and E-Mail Address. This information will not be published and is for contact purposes only.

Authors should submit their work via E-mail to the Indiana Undergraduate Journal of Cognitive Science Editorial Board at iacs@indiana.edu. Once your submission is received, a confirmation E-mail will be sent by the Editorial Board. All submissions will be considered equally and no preference will be given to any particular discipline within cognitive science.

III. Review / Acceptance Process

After submission, the Editorial Board and a panel of reviewers will review all articles and will decide which papers will be published in the current edition of the journal. Authors will be notified by E-mail if their paper is accepted for publication. Authors will also be notified by E-mail if their paper is not selected for publication. After acceptance, authors are required to submit their paper in Microsoft Word format to allow the Editorial Board to make formatting and grammatical edits. Once these changes have been made, the Editorial Board will contact the authors to obtain final approval of the above changes and to obtain written publication permission.

IV. Disclaimer and More Information

Articles published in the Indiana Undergraduate Journal of Cognitive Science are considered copyrighted. However, this journal is not a binding publication. Authors are free to submit their work to any other publications they wish.

For more information about the journal, please contact Michael Amlung, Executive Editor of the Indiana Undergraduate Journal of Cognitive Science, at mamlung@indiana.edu or by the means below.

Indiana Undergraduate Journal of Cognitive Science
819 Eigenmann Hall - 1910 E. 10th St.
Indiana University - Bloomington, Indiana 47406
E-Mail: iacs@indiana.edu
<http://www.cogs.indiana.edu/iacs>

