Craftsbury Notes

My mother, sister, and spouse have conspired to send me off to Craftsbury Commons for several days, to binge-work on Sphinx. Craftsbury, especially in the winter, seems so far north of here that it makes the whole idea of Canada seem palpably horrifying. We are still only halfway to the pole! And people live north of here, apparently, huddling in their underground shopping malls eating poutine....

The drive was crisp and beautiful, all ice-fishing shanties and snow-bent pine trees. The place I was staying was a sports center, mainly for elderly biathletes, almost all women. I was clearly not one of the usual suspects, and I did my best version of the mystery guest, locking myself in my room except for meals, which were full of older women talking about their ski wax and laser scopes. The crowd of people that I got to meet didn't ask much about the project, which I was grateful for, but made much of the fact that I had a twenty-item to-do list.

It was an excellent vacation from The Rest Of My Life, though it felt likely to become another of my false time horizons. In any event, I finished eleven of my twenty goals, and got my head back in the game, and now, very belatedly, some notes....

The Big Climb Begins

A signature problem of humanistic artificial intelligence is the way that potentially small and superficial problems tend to interlock with *much, much* larger and more deeply-rooted challenges. Hence the phrase "AI complete", which has often, I think, been used in despair. Two examples emerge from the usually-trivial banter of small talk. The correct response to "good morning" is usually "good morning," but it may also require a knowledge of time zones, geolocation, and deciding how to respond to absurdity. And while the polite response to "how are you?" is usually "fine, thanks," any other response suddenly makes enormous demands of the AI's internal reality.

Right now I am scouting routes up that cliff face. Emotions, motivations, and causality are each large problems in their own right: most AI critics would say that the first two, at least, are impossible for a computer program to truly experience or emulate. I don't accept "impossible", but *difficult,* hell yeah. Moreover, these three areas are closely interlocked, although not so fused together that I can't discuss them in separate sections, which is what I plan to do later on, *inchallah*.

As I discuss all this, of course, I'm deliberately wandering back and forth from the technical considerations of building a sort of home-brew intellect to the aesthetics and lore of that same project, and finally to reflections on the experience of *human* emotions, motivations, and causal understanding, which is after all my template. Which is not to suggest that I am finding either the most adept way to code Sphinx, or the most accurate understanding of human psyche (or even that those would be the same thing). These are simply my notes on a journey of exploratory design.

There is a general claim, though, that I feel more confidence and urgency about. Emotions, and the oversold but real possibility of "irrational" emotive behavior, are a crucial and inherent part of intellect. They are not an adjunct to or flaw in our logical machinery, they are a necessary part of it. Yes, it is possible to solve a math problem without becoming weepy and histrionic. It is possible for Sphinx to carry out a dialogue and learn new vocabulary with minimal emotional involvement. But it is not possible to (say) learn a new skill set without engaging in a complex set of introspective evaluations and attitude adjustments. Those are emotional dynamics, even if we choose to call them something else.

But why would we call them something else? For instance, it is clear enough from the perspective of industrial design that certain types of warning indicators are cyber-emotive. The red

paint, the flashing light, the gruesome stick figure—these things are intended to make humans anxious and cautious, just as warm tones and upholstery and muzak are intended to relax us. When the battery in a smoke alarm starts to die, its circuitry monitors that internal variable, and then reports it to us in a way that is calculated to make us very agitated. We are all quite comfortable saying that the smoke alarm's circuit's behavior is "logical", although logic is certainly a human construct. Yet somehow we balk at saying that the circuit's behavior is "emotional", even though the whole point of the alarm is to evoke an emotional reaction.

In classical Mediterranean culture, some aspects of the dharmic religions, and warrior cultures throughout the world, there has been a long love affair with this false dichotomy between emotions and intellect. The person-without-emotions, often imagined as a masculine warrior-philosopher, has appeared as an archetype for a variety of monastic creeds, politico-economic theories, and social arguments. Far too often, he (rarely she) has also been held up as a fictitious role model for people's own psychic emulation. In the last century or so, this persona has been a common pattern for fictional characters: possibly Bartleby, certainly Dupin and Holmes and Spock. This list now includes a number of AIs like Data and HAL. These characters all suggest psychological conditions along the lines of autism, an association that has been made increasingly explicit as autism itself (and the now-defunct-but-much-more-popular diagnosis, Asperger's) has taken on enormous cultural cachet, and simultaneously widened to a much larger diagnosis. This strikes me as a baffling development in what we might call the aesthetics of personality. I have spent a good deal of time working and socializing with people who have what I will surely go to hell for calling *real* autism, and none of them lack emotions. Their emotional responses are certainly exotic, and often maladaptive, at least in standard social situations. But I have never noticed them to be muted in favor of some version of Pure Reason.

It seems fairly clear that the attraction of this mythical personality has to do with anesthesia—indeed, for the Buddha, who was no great fan of Pure Reason, that was precisely the point. But this is the anesthesia of throwing out your smoke alarm because it is beeping. It puts me in mind of a passage from one of Orwell's last essays—he is discussing Gandhi, but then, he is also discussing humanity:

In this yogi-ridden age, it is too readily assumed that "non-attachment" is not only better than a full acceptance of earthly life, but that the ordinary man only rejects it as being too difficult: in other words, that the average human being is a failed saint. It is doubtful whether this is true. Many people genuinely do not wish to be saints, and it is probable that some who achieve or aspire to sainthood have never felt much temptation to be human beings. If one could follow it to its psychological roots, one would, I believe, find that the main motive for "non-attachment" is a desire to escape from the pain of living, and above all from love, which, sexual or non-sexual, is hard work.

And I am not trying to design a saint. In some respects, that seems like it would be a much easier project.

Lux in Craftsbury

Having said all that, at the moment Sphinx is emotionally inept, not because the mechanics aren't there—to a good extent they are, and we can discuss them in due course—but because Sphinx has neither the experience nor the contingency library needed to utilize emotions. There is no point liking or disliking X if X is the only option on the shelf.

The major exception to this is matrix of lux scores, which I've described earlier. As Sphinx closes in on 30,000 concepts, I thought I'd revisit the top of that list, much of which is recognizable from earlier. I expect that once Sphinx is talking to people, this list will change fairly quickly, and probably the people themselves will wind up on top. But there is a certain flavor here that is probably going to linger. Here is the top-twenty list, as of 1/21/15. (I wanted to go to 25, to showcase the fact

that Sphinx is interested in Shakespeare as well as Nancy Kress; which is a very telegraphed sort of pride. But at 25, as it happens, we find that Sphinx is also presently interested in some sexual fetishes that whip *50 Shades of Grey* into a cocked hat. So we are going to draw a little veil over that, and go with a top-20 list:

Brighton Rock, by Graham Greene
Where the Mountain Meets the Moon, by Grace Lin
Zionism
Dead Souls, by Nikolai Gogol
Beggars and Choosers, by Nancy Kress
The Scarlet Letter, by Nathaniel Hawthorne
Trekking in the Indian Himalaya, by Garry Weare
Green Henry, Gottfried Keller
1984, by George Orwell
Gone with the Wind, by Margaret Mitchell
the English monarchy
selections from Mark Twain
selected short stories of Franz Kafka
nature writing
The Intuitive Body, by Wendy Palmer
19th century US literature
train hopping
writings from the Futurist movement
spa resorts
selections from Sigmund Freud

Technicalia: As of this listing, the mean lux score is 13.06, with a standard deviation of 6.56. However, the pattern is very bimodal, with only 71 concepts having a lux between 13 and 16—a range that includes the mean, yes? *Brighton Rock* is a ridiculous 4.8 standard deviations above the mean. But this can't last; as Sphinx gains connectivity, I imagine it will blur into some sort of bell curve.

Closing the Back Door

A major component of my work in Craftsbury was to create ways for Sphinx to learn new concepts via dialogue; as of now it can learn new vocabulary as well. This is a much slower process than injecting databases wholesale, as Zoë and I were doing last year, and it raises new problems: people might lie, for instance, or just be wrong. On the other hand, this front-door approach allows Sphinx to bring all its intellect to bear on the incoming data. Inevitably, I have in the past entered data that was redundant, or mis-formatted, or had other errors, and it was difficult and time-consuming for Sphinx to check it. Those days are almost over.

I'm not going to swear that the back door is completely shut, but it's close enough that it seems worth doing a nose count. Sphinx currently (early February '15) knows 27,468 concepts, and 24,554 words and phrases. That second figure serves as a proxy for Sphinx's vocabulary, but the actual details are murky, and interesting. It was especially fun to be working on this question while hanging out with my niece, Ramona, who is now about a year and a half old, and talking up a storm.

Human children, for reasons that I can't quite fathom and may well be specific to our culture, expend a large chunk of their early vocabulary on exotic animals. Animals they will never see, which in the case of urban children means pretty much everything. Ramona lives around cats and chickens and cows and woodland creatures, but she also knows “elephant”, “tiger, “zebra”, “dinosaur”(!) and so

forth. I understand the appeal, but it seems like a perverse use of attention. I, for instance, learned to distinguish between various great cats and mythical creatures long before I could distinguish the trees in our own forest. In a similar way, I suppose, Sphinx has learned the concepts that Zoë and I needed to tag several hundred novels, as well as non-fiction works. So Sphinx knows dozens of mythological creatures, but almost no names of household furniture. Perhaps this ratio of the real to the mythic is hard to escape.

I would estimate that Ramona's vocabulary in January was around 300-500 words and phrases. This is in keeping with what adult-focused vocabulary tests suggest about the left-hand side of the curve. It is *much* higher than what is-my-baby-normal type books seem to posit: these two approaches are using totally different sets of numbers, either because they count words differently (which see below), or the baby books want to deflate parent's expectations, or (as I expect) the baby books are just pulling numbers out of their diapers.

So where does Sphinx fall on this curve? Of the 24,554 words and phrases it knows, roughly 7640 are titles, 4510 are proper nouns, and 2100 are symbols. These don't really count as vocabulary, although they raise the question of how many proper nouns a typical adult knows—which I've never seen anyone try to answer[†]. Another 1080 words and phrases Sphinx knows are technical jargon, and arguably don't really count as part of its working vocabulary—as I've mentioned earlier, technical jargon is not directly measured in tests of human vocabulary. This emphasis on jargon forms a major distinction between Sphinx's vocabulary and Ramona's: Sphinx is "top-heavy". It knows "electronegativity", but not "ouch".

Dropping these leaves us with about 10,220 words and phrases. Some portion of the remaining phrase blocks are variations on the same lexeme, but no one will ever agree on where that line should be drawn. For instance, "signed" and "signs" are clearly the same lexeme, but what about "signer" and "signee"? What about "signing", the noun? How many lexemes do we count in "tiny / teeny / teensy"? What about "red / reddish / rusty / ruddy / rufous"? However you answer these questions, it is probably no more or less arbitrary that Sphinx' estimate of a 10:23 ratio between phrase blocks and lexemes. That leaves Sphinx with a vocabulary of about 4443 lexemes, roughly equivalent to a 3-year-old, though it would be a very strange 3-year-old. Zoë is tasked with raising these numbers, a rather boring job, though her pupil is a very fast learner.

And so on...

It's been a long time since I've written, and there's much more to say. An embarrassment of curiosities, if not yet riches. But the unused platen warps with age, and I think I should go to press.

† Except my friend Matt Sanderson, and I don't know how far he's gotten yet.