

Concept Safety

1. [Concept Safety: Producing similar AI-human concept spaces](#)
2. [Concept Safety: The problem of alien concepts](#)
3. [Concept Safety: What are concepts for, and how to deal with alien concepts](#)
4. [Concept Safety: World-models as tools](#)

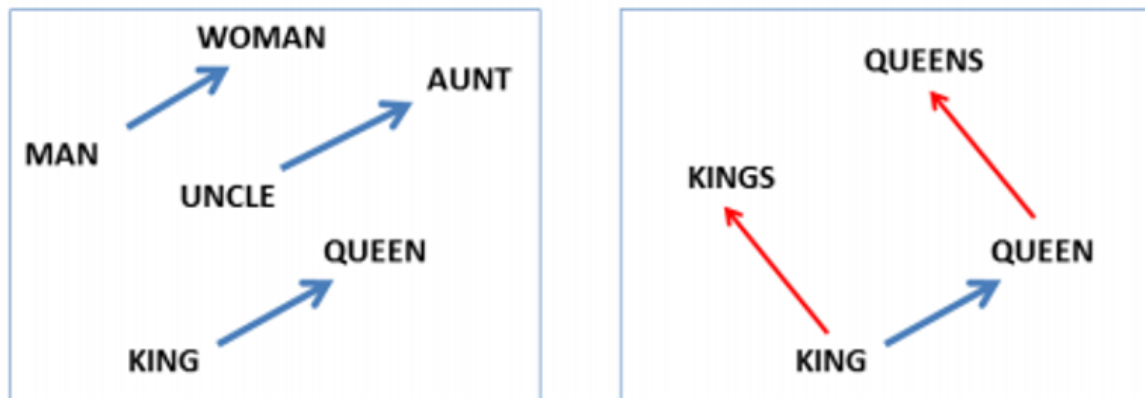
Concept Safety: Producing similar AI-human concept spaces

A frequently-raised worry about AI is that it may reason in ways which are very different from us, and understand the world in a very alien manner. For example, [Armstrong, Sandberg & Bostrom \(2012\)](#) consider the possibility of restricting an AI via "rule-based motivational control" and programming it to follow restrictions like "stay within this lead box here", but they raise worries about the difficulty of rigorously defining "this lead box here". To address this, they go on to consider the possibility of making an AI internalize human concepts via feedback, with the AI being told whether or not some behavior is good or bad and then constructing a corresponding world-model based on that. The authors are however worried that this may fail, because

Humans seem quite adept at constructing the correct generalisations – most of us have correctly deduced what we should/should not be doing in general situations (whether or not we follow those rules). But humans share a common of genetic design, which the OAI would likely not have. Sharing, for instance, derives partially from genetic predisposition to reciprocal altruism: the OAI may not integrate the same concept as a human child would. Though reinforcement learning has a good track record, it is neither a panacea nor a guarantee that the OAIs generalisations agree with ours.

Addressing this, a possibility that I raised in [Sotola \(2015\)](#) was that possibly the concept-learning mechanisms in the human brain are actually relatively simple, and that we could replicate the human concept learning process by replicating those rules. I'll start this post by discussing a closely related hypothesis: that *given a specific learning or reasoning task and a certain kind of data, there is an optimal way to organize the data that will naturally emerge*. If this were the case, then AI and human reasoning might naturally tend to learn the same kinds of concepts, even if they were using very different mechanisms. Later on the post, I will discuss how one might try to verify that similar representations had in fact been learned, and how to set up a system to make them even more similar.

Word embedding



A particularly fascinating branch of recent research relates to the learning of word embeddings, which are mappings of words to very high-dimensional vectors. It turns out that if you train a system on one of several kinds of tasks, such as being able to classify sentences as valid or invalid, this builds up a space of word vectors that reflects the relationships between the words. For example, there seems to be a male/female dimension to words, so that there's a "female vector" that we can add to the word "man" to get "woman" - or, equivalently, which we can subtract from "woman" to get "man". And it so

happens ([Mikolov, Yih & Zweig 2013](#)) that we can also get from the word "king" to the word "queen" by adding the same vector to "king". In general, we can (roughly) get to the male/female version of any word vector by adding or subtracting this one difference vector!

Why would this happen? Well, a learner that needs to classify sentences as valid or invalid needs to classify the sentence "the king sat on his throne" as valid while classifying the sentence "the king sat on her throne" as invalid. So including a gender dimension on the built-up representation makes sense.

But gender isn't the only kind of relationship that gets reflected in the geometry of the word space. Here are a few more:

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

It turns out ([Mikolov et al. 2013](#)) that with the right kind of training mechanism, a *lot* of relationships that we're intuitively aware of become automatically learned and represented in the concept geometry. And like [Olah \(2014\)](#) comments:

It's important to appreciate that all of these properties of W are *side effects*. We didn't try to have similar words be close together. We didn't try to have analogies encoded with difference vectors. All we tried to do was perform a simple task, like predicting whether a sentence was valid. These properties more or less popped out of the optimization process.

This seems to be a great strength of neural networks: they learn better ways to represent data, automatically. Representing data well, in turn, seems to be essential to success at many machine learning problems. Word embeddings are just a particularly striking example of learning a representation.

It gets even more interesting, for we can use these for translation. Since Olah has already written an excellent exposition of this, I'll just quote him:

We can learn to embed words from two different languages in a single, shared space. In this case, we learn to embed English and Mandarin Chinese words in the same space.

We train two word embeddings, W_{en} and W_{zh} in a manner similar to how we did above. However, we know that certain English words and Chinese words have similar meanings. So, we optimize for an additional property: words that we know are close translations should be close together.

Of course, we observe that the words we knew had similar meanings end up close together. Since we optimized for that, it's not surprising. More interesting is that words we *didn't* know were translations end up close together.

In light of our previous experiences with word embeddings, this may not seem too surprising. Word embeddings pull similar words together, so if an English and Chinese word we know to mean similar things are near each other, their synonyms will also end up near each other. We also know that things like gender differences tend to end up being represented with a constant difference vector. It seems like forcing enough points to line up should force these difference vectors to be the same in both the English and Chinese embeddings. A result of this would be that if we know that two male versions of words translate to each other, we should also get the female words to translate to each other.

Intuitively, it feels a bit like the two languages have a similar 'shape' and that by forcing them to line up at different points, they overlap and other points get pulled into the right positions.

After this, it gets *even more interesting*. Suppose you had this space of word vectors, and then you also had a system which translated *images* into vectors in the same space. If you have images of dogs, you put them near the word vector for dog. If you have images of Clippy you put them near word vector for "paperclip". And so on.

You do that, and then you take some class of images the image-classifier was never trained on, like images of cats. You ask it to place the cat-image somewhere in the vector space. Where does it end up?

You guessed it: in the rough region of the "cat" words. Olah once more:

This was done by members of the Stanford group with only 8 known classes (and 2 unknown classes). The results are already quite impressive. But with so few known classes, there are very few points to interpolate the relationship between images and semantic space off of.

The Google group did a much larger version – instead of 8 categories, they used 1,000 – around the same time ([Frome et al. \(2013\)](#)) and has followed up with a new variation ([Norouzi et al. \(2014\)](#)). Both are based on a very powerful image classification model (from [Krizhevsky et al. \(2012\)](#)), but embed images into the word embedding space in different ways.

The results are impressive. While they may not get images of unknown classes to the precise vector representing that class, they are able to get to the right neighborhood. So, if you ask it to classify images of unknown classes and the classes are fairly different, it can distinguish between the different classes.

Even though I've never seen a Aesculapian snake or an Armadillo before, if you show me a picture of one and a picture of the other, I can tell you which is which because I have a general idea of what sort of animal is associated with each word. These networks can accomplish the same thing.

These algorithms made no attempt of being biologically realistic in any way. They didn't try classifying data the way the brain does it: they just tried classifying data using whatever worked. And it turned out that this was enough to start constructing a multimodal representation space where a lot of the relationships between entities were similar to the way humans understand the world.

How useful is this?

"Well, that's cool", you might now say. "But those word spaces were constructed from human linguistic data, for the purpose of predicting human sentences. Of course they're going to classify the world in the same way as humans do: they're basically learning the human representation of the world. That doesn't mean that an autonomously learning AI, with its

own learning faculties and systems, is necessarily going to learn a similar internal representation, or to have similar concepts."

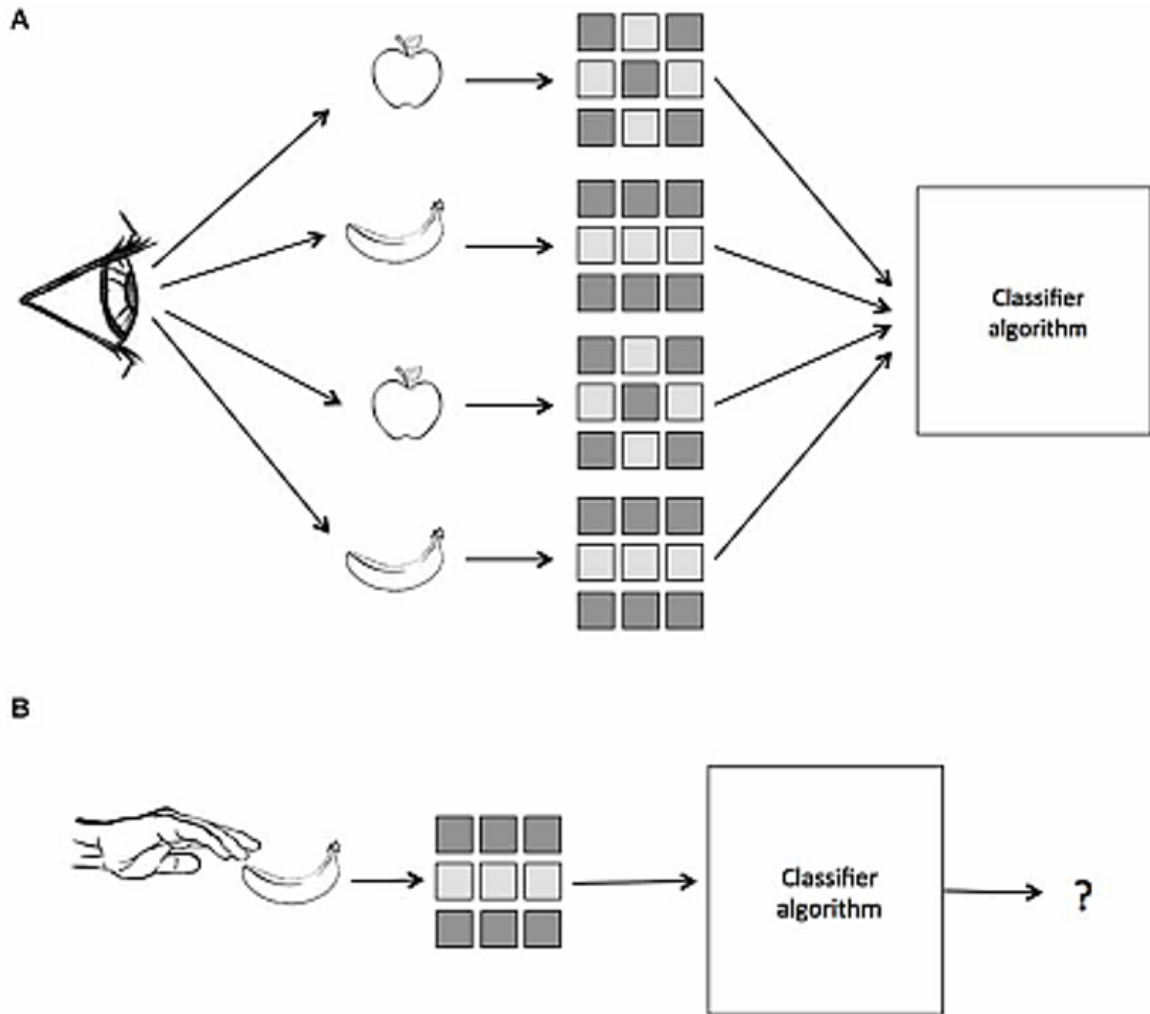
This is a fair criticism. But it is *mildly suggestive* of the possibility that an AI that was trained to understand the world via feedback from human operators would end up building a similar conceptual space. At least assuming that we chose the right learning algorithms.

When we train a language model to classify sentences by labeling some of them as valid and others as invalid, there's a hidden structure implicit in our answers: the structure of how we understand the world, and of how we think of the meaning of words. The language model extracts that hidden structure and begins to classify previously unseen things in terms of those implicit reasoning patterns. Similarly, if we gave an AI feedback about what kinds of actions counted as "leaving the box" and which ones didn't, there would be a certain way of viewing and conceptualizing the world implied by that feedback, one which the AI could learn.

Comparing representations

"Hmm, *maaaaaaaaybe*", is your skeptical answer. "But how would you ever know? Like, you can test the AI in your training situation, but how do you know that it's actually acquired a similar-enough representation and not something wildly off? And it's one thing to look at those vector spaces and *claim* that there are human-like relationships among the different items, but that's still a little hand-wavy. We don't actually *know* that the human brain does anything remotely similar to represent concepts."

Here we turn, for a moment, to neuroscience.



[Multivariate Cross-Classification](#) (MVCC) is a clever neuroscience methodology used for figuring out whether different neural representations of the same thing have something in common. For example, we may be interested in whether the visual and tactile representation of a banana have something in common.

We can test this by having several test subjects look at pictures of objects such as apples and bananas while sitting in a brain scanner. We then feed the scans of their brains into a machine learning classifier and teach it to distinguish between the neural activity of looking at an apple, versus the neural activity of looking at a banana. Next we have our test subjects (still sitting in the brain scanners) *touch* some bananas and apples, and ask our machine learning classifier to guess whether the resulting neural activity is the result of touching a banana or an apple. If the classifier - which has *not* been trained on the "touch" representations, only on the "sight" representations - manages to achieve a better-than-chance performance on this latter task, then we can conclude that the neural representation for e.g. "the sight of a banana" has something in common with the neural representation for "the touch of a banana".

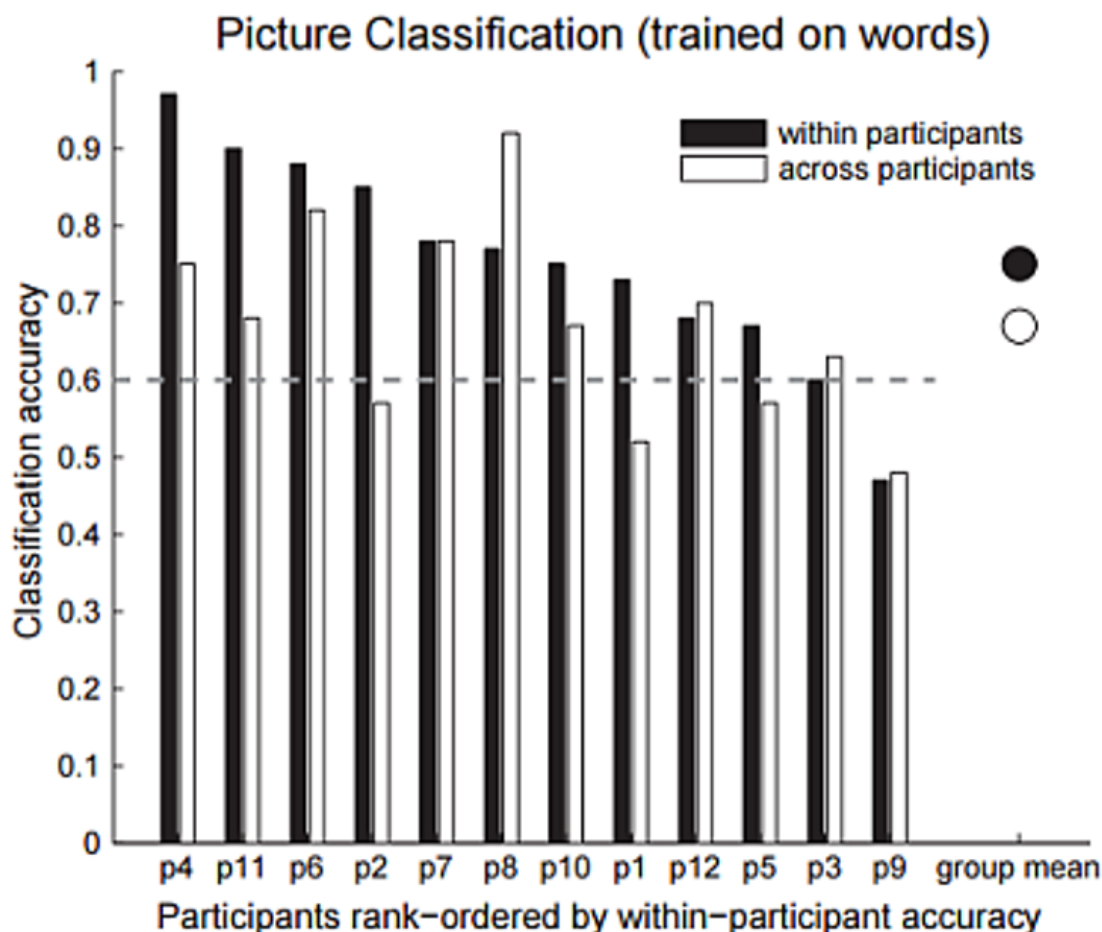


Fig. 2. Classification accuracies across stimulus formats when training on words to identify the category of viewed pictures. Reliable ($p < 0.05$) accuracies for identification of category of viewed pictures (filled bars) were reached for 11 participants when training on word data, and reliable ($p < 0.05$) accuracies for identification of category of viewed pictures when training on the union of word data from the other participants (unfilled bars) were reached for 8 out of 12 participants. The dashed line indicates the $\alpha = 0.05$ level of significance.

A particularly fascinating experiment of this type is that of [Shinkareva et al. \(2011\)](#), who showed their test subjects both the written words for different tools and dwellings, and, separately, line-drawing images of the same tools and dwellings. A machine-learning classifier was both trained on image-evoked activity and made to predict word-evoked activity and vice versa, and achieved a high accuracy on category classification for both tasks. Even more interestingly, the representations seemed to be similar between subjects. Training the classifier on the word representations of all but one participant, and then having it classify the image representation of the left-out participant, also achieved a reliable ($p < 0.05$) category classification for 8 out of 12 participants. This suggests a relatively similar concept space between humans of a similar background.

We can now hypothesize some ways of testing the similarity of the AI's concept space with that of humans. Possibly the most interesting one might be to develop a translation between a human's and an AI's internal representations of concepts. Take a human's neural activation

when they're thinking of some concept, and then take the AI's internal activation when it is thinking of the same concept, and plot them in a shared space similar to the English-Mandarin translation. To what extent do the two concept geometries have similar shapes, allowing one to take a human's neural activation of the word "cat" to find the AI's internal representation of the word "cat"? To the extent that this is possible, one could probably establish that the two share highly similar concept systems.

One could also try to more explicitly optimize for such a similarity. For instance, one could train the AI to make predictions of different concepts, with the additional constraint that its internal representation must be such that a machine-learning classifier trained on a human's neural representations will correctly identify concept-clusters within the AI. This might force internal similarities on the representation beyond the ones that would already be formed from similarities in the data.

Next post in series: [The problem of alien concepts](#).

Concept Safety: The problem of alien concepts

In the [previous post](#) in this series, I talked about how one might get an AI to have similar concepts as humans. However, one would intuitively assume that a superintelligent AI might eventually develop the capability to entertain far more sophisticated concepts than humans would ever be capable of having. Is that a problem?

Just what are concepts, anyway?

To answer the question, we first need to define what exactly it is that we mean by a "concept", and why exactly more sophisticated concepts would be a problem.

Unfortunately, there isn't really any standard definition of this in the literature, with different theorists having different definitions. [Machery](#) even argues that the term "concept" doesn't refer to a natural kind, and that we should just get rid of the whole term. If nothing else, this definition from [Kruschke \(2008\)](#) is at least amusing:

Models of categorization are usually designed to address data from laboratory experiments, so "categorization" might be best defined as the class of behavioral data generated by experiments that ostensibly study categorization.

Because I don't really have the time to survey the whole literature and try to come up with one grand theory of the subject, I will for now limit my scope and only consider two compatible definitions of the term.

Definition 1: Concepts as multimodal neural representations. I touched upon this definition in the last post, where I mentioned studies indicating that the brain seems to have shared neural representations for e.g. the touch and sight of a banana. Current neuroscience seems to indicate the existence of brain areas where representations from several different senses are combined together into higher-level representations, and where the activation of any such higher-level representation will also end up activating the lower sense modalities in turn. As summarized by [Man et al. \(2013\)](#):

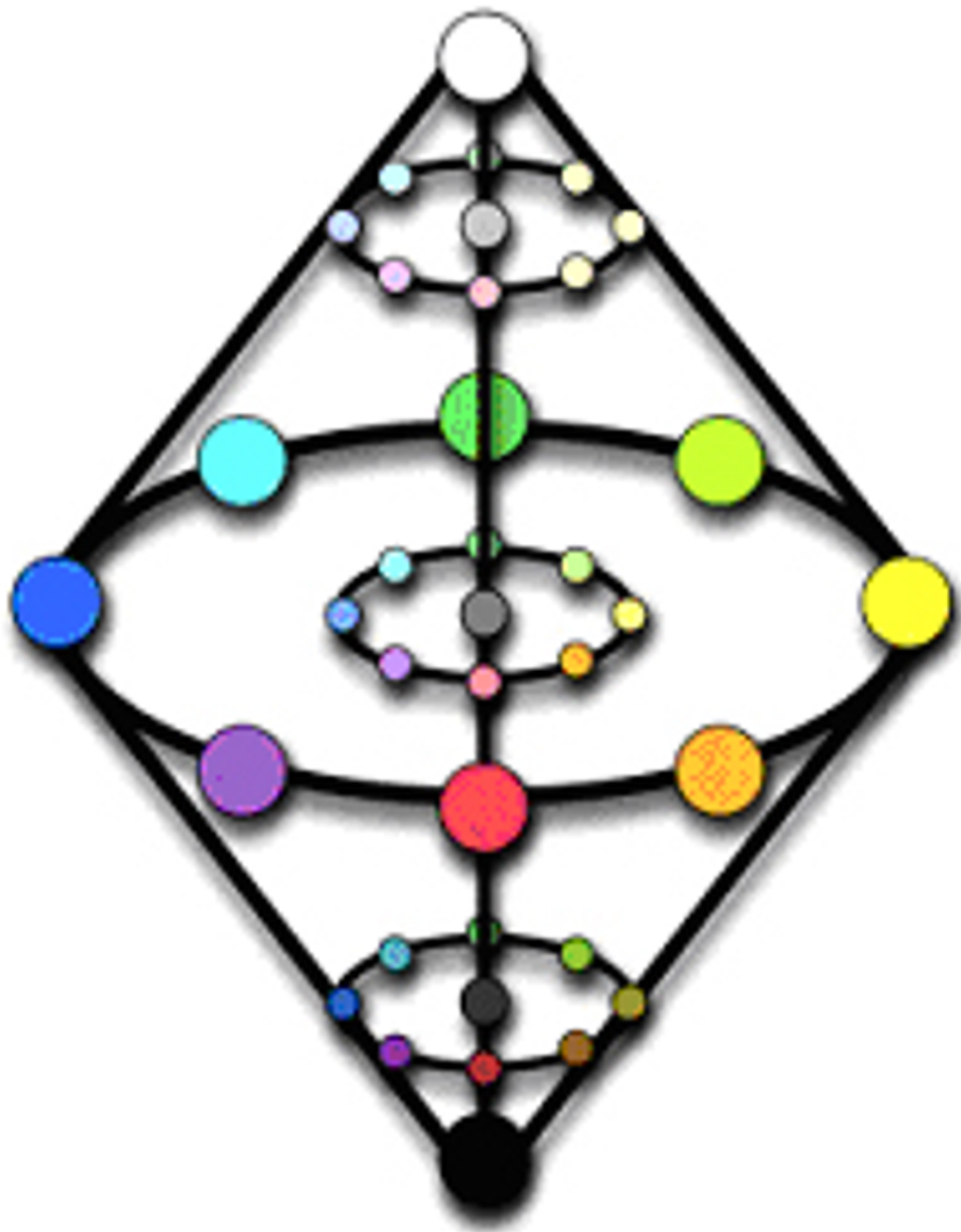
Briefly, the Damasio framework proposes an architecture of convergence-divergence zones (CDZ) and a mechanism of time-locked retroactivation. Convergence-divergence zones are arranged in a multi-level hierarchy, with higher-level CDZs being both sensitive to, and capable of reinstating, specific patterns of activity in lower-level CDZs. Successive levels of CDZs are tuned to detect increasingly complex features. Each more-complex feature is defined by the conjunction and configuration of multiple less-complex features detected by the preceding level. CDZs at the highest levels of the hierarchy achieve the highest level of semantic and contextual integration, across all sensory modalities. At the foundations of the hierarchy lie the early sensory cortices, each containing a mapped (i.e., retinotopic, tonotopic, or somatotopic) representation of sensory space. When a CDZ is activated by an input pattern that resembles the template for which it has been tuned, it retro-activates the template pattern of lower-level CDZs. This continues down the hierarchy of CDZs, resulting in an ensemble of well-specified and time-locked activity extending to the early sensory cortices.

On this account, my mental concept for "dog" consists of a neural activation pattern making up the sight, sound, etc. of some dog - either a generic prototypical dog or some more specific dog. Likely the pattern is not just limited to sensory information, either, but may be associated with e.g. motor programs related to dogs. For example, the program for throwing a ball for the dog to fetch. One version of this hypothesis, the Perceptual Symbol Systems account, calls such multimodal representations *simulators*, and describes them as follows ([Niedenthal et al. 2005](#)):

A simulator integrates the modality-specific content of a category across instances and provides the ability to identify items encountered subsequently as instances of the same category. Consider a simulator for the social category, politician. Following exposure to different politicians, visual information about how typical politicians look (i.e., based on their typical age, sex, and role constraints on their dress and their facial expressions) becomes integrated in the simulator, along with auditory information for how they typically sound when they talk (or scream or grovel), motor programs for interacting with them, typical emotional responses induced in interactions or exposures to them, and so forth. The consequence is a system distributed throughout the brain's feature and association areas that essentially represents knowledge of the social category, politician.

The inclusion of such "extra-sensory" features helps understand how even abstract concepts could fit this framework: for example, one's understanding of the concept of a derivative might be partially linked to the procedural programs one has developed while solving derivatives. For a more detailed hypothesis of how abstract mathematics may emerge from basic sensory and motor programs and concepts, I recommend [Lakoff & Nuñez \(2001\)](#).

Definition 2: Concepts as areas in a psychological space. This definition, while being compatible with the previous one, looks at concepts more "from the inside". [Gärdenfors \(2000\)](#) defines the basic building blocks of a psychological conceptual space to be various *quality dimensions*, such as temperature, weight, brightness, pitch, and the spatial dimensions of height, width, and depth. These are *psychological* in the sense of being derived from our phenomenal experience of certain kinds of properties, rather than the way in which they might exist in some objective reality.



For example, one way of modeling the psychological sense of color is via a color space defined by the quality dimensions of hue (represented by the familiar color circle), chromaticness (saturation), and brightness.

The second phenomenal dimension of color is *chromaticness* (saturation), which ranges from grey (zero color intensity) to increasingly greater intensities. This dimension is isomorphic to an interval of the real line. The third dimension is *brightness* which varies from white to black and is thus a linear dimension with two end points. The two latter dimensions are not totally independent, since the possible variation of the chromaticness dimension decreases as the values of the brightness dimension approaches the extreme

points of black and white, respectively. In other words, for an almost white or almost black color, there can be very little variation in its chromaticness. This is modeled by letting that chromaticness and brightness dimension together generate a triangular representation ... Together these three dimensions, one with circular structure and two with linear, make up the color space. This space is often illustrated by the so called color spindle

This kind of a representation is different from the physical wavelength representation of color, where e.g. the hue is mostly related to the wavelength of the color. The wavelength representation of hue would be linear, but due to the properties of the human visual system, the psychological representation of hue is circular.

Gärdenfors defines two quality dimensions to be *integral* if a value cannot be given for an object on one dimension without also giving it a value for the other dimension: for example, an object cannot be given a hue value without also giving it a brightness value. Dimensions that are not integral with each other are *separable*. A *conceptual domain* is a set of integral dimensions that are separable from all other dimensions: for example, the three color-dimensions form the domain of color.

From these definitions, Gärdenfors develops a theory of concepts where more complicated conceptual spaces can be formed by combining lower-level domains. Concepts, then, are particular regions in these conceptual spaces: for example, the concept of "blue" can be defined as a particular region in the domain of color. Notice that the notion of various combinations of basic perceptual domains making more complicated conceptual spaces possible fits well together with the models discussed in our previous definition. There more complicated concepts were made possible by combining basic neural representations for e.g. different sensory modalities.

The origin of the different quality dimensions could also emerge from the specific properties of the different simulators, as in PSS theory.

Thus definition #1 allows us to talk about what a concept might "look like from the outside", with definition #2 talking about what the same concept might "look like from the inside".

Interestingly, Gärdenfors hypothesizes that much of the work involved with learning new concepts has to do with learning new quality dimensions to fit into one's conceptual space, and that once this is done, all that remains is the comparatively much simpler task of just dividing up the new domain to match seen examples.

For example, consider the (phenomenal) dimension of volume. The experiments on "conservation" performed by Piaget and his followers indicate that small children have no separate representation of volume; they confuse the volume of a liquid with the height of the liquid in its container. It is only at about the age of five years that they learn to represent the two dimensions separately. Similarly, three- and four-year-olds confuse high with tall, big with bright, and so forth (Carey 1978).

The problem of alien concepts

With these definitions for concepts, we can now consider what problems would follow if we started off with a very human-like AI that had the same concepts as we did, but then expanded its conceptual space to allow for entirely new kinds of concepts. This could happen if it self-modified to have new kinds of sensory or thought modalities that it could associate its existing concepts with, thus developing new kinds of quality dimensions.

An analogy helps demonstrate this problem: suppose that you're operating in a two-dimensional space, where a rectangle has been drawn to mark a certain area as "forbidden" or "allowed". Say that you're an inhabitant of [Flatland](#). But then you suddenly become aware that actually, the world is three-dimensional, and has a height dimension as well! That raises the question of, how should the "forbidden" or "allowed" area be understood in this new

three-dimensional world? Do the walls of the rectangle extend infinitely in the height dimension, or perhaps just some certain distance in it? If just a certain distance, does the rectangle have a "roof" or "floor", or can you just enter (or leave) the rectangle from the top or the bottom? There doesn't seem to be any clear way to tell.

As a historical curiosity, this dilemma actually kind of really happened when airplanes were invented: could landowners forbid airplanes from flying over their land, or was the ownership of the land limited to some specific height, above which the landowners had no control? Courts and legislation eventually settled on the latter answer. A more AI-relevant example might be if one was trying to limit the AI with rules such as "stay within this box here", and the AI then gained an intuitive understanding of quantum mechanics, which might allow it to escape from the box without violating the rule in terms of its new concept space.

More generally, if previously your concepts had N dimensions and now they have $N+1$, you might find something that fulfilled all the previous criteria while still being different from what we'd prefer if we knew about the $N+1$ th dimension.

In the next post, I will present some (very preliminary and probably wrong) ideas for solving this problem.

Next post in series: [What are concepts for, and how to deal with alien concepts.](#)

Concept Safety: What are concepts for, and how to deal with alien concepts

In [The Problem of Alien Concepts](#), I posed the following question: if your concepts (defined as either multimodal representations or as areas in a psychological space) previously had N dimensions and then they suddenly have $N+1$, how does that affect (moral) values that were previously only defined in terms of N dimensions?

I gave some (more or less) concrete examples of this kind of a "conceptual expansion":

1. Children learn to represent dimensions such as "height" and "volume", as well as "big" and "bright", separately at around age 5.
2. As an inhabitant of the Earth, you've been used to people being unable to fly and landowners being able to forbid others from using their land. Then someone goes and invents an airplane, leaving open the question of the height to which the landowner's control extends. Similarly for [satellites and nation-states](#).
3. As an inhabitant of Flatland, you've been told that the inside of a certain rectangle is a forbidden territory. Then you learn that the world is actually three-dimensional, leaving open the question of the height of which the forbidden territory extends.
4. An AI has previously been reasoning in terms of classical physics and been told that it can't leave a box, which it previously defined in terms of classical physics. Then it learns about quantum physics, which allow for definitions of "location" which are substantially different from the classical ones.

As a hint of the direction where I'll be going, let's first take a look at how *humans* solve these kinds of dilemmas, and consider examples #1 and #2.

The first example - children realizing that items have a volume that's separate from their height - rarely causes any particular crises. Few children have values that would be seriously undermined or otherwise affected by this discovery. We might say that *it's a non-issue because none of the children's values have been defined in terms of the affected conceptual domain*.

As for the second example, I don't know the exact cognitive process by which it was decided that you didn't need the landowner's permission to fly over their land. But I'm guessing that it involved reasoning like: if the plane flies at a sufficient height, then that doesn't harm the landowner in any way. Flying would become impossible difficult if you had to get separate permission from every person whose land you were going to fly over. And, especially before the invention of radar, a ban on unauthorized flyovers would be next to impossible to enforce anyway.

We might say that after an option became available which forced us to include a new dimension in our existing concept of landownership, we *solved the issue by considering it in terms of our existing values*.

Concepts, values, and reinforcement learning

Before we go on, we need to talk a bit about why we have concepts and values in the first place.

From an evolutionary perspective, creatures that are better capable of harvesting resources (such as food and mates) and avoiding dangers (such as other creatures who think you're food or after their mates) tend to survive and have offspring at better rates than otherwise comparable creatures who are worse at those things. If a creature is to be flexible and capable of responding to novel situations, it can't just have a pre-programmed set of responses to different things. Instead, it needs to be able to learn how to harvest resources and avoid danger even when things are different from before.

How did evolution achieve that? Essentially, by creating a brain architecture that can, as a very very rough approximation, be seen as consisting of two different parts. One part, which a machine learning researcher might call the *reward function*, has the task of figuring out when various criteria - such as being hungry or getting food - are met, and issuing the rest of the system either a positive or negative reward based on those conditions. The other part, the *learner*, then "only" needs to find out how to best optimize for the maximum reward. (And then there is the third part, which includes any region of the brain that's neither of the above, but we don't care about those regions now.)

The mathematical theory of how to learn to optimize for rewards when your environment and reward function are unknown is reinforcement learning (RL), which recent neuroscience indicates [is implemented by the brain](#). An RL agent learns a mapping from states of the world to rewards, as well as a mapping from actions to world-states, and then uses that information to maximize the amount of lifetime rewards it will get.

There are two major reasons why an RL agent, like a human, should learn high-level *concepts*:

1. They make learning massively easier. Instead of having to separately learn that "in the world-state where I'm sitting naked in my cave and have berries in my hand, putting them in my mouth enables me to eat them" and that "in the world-state where I'm standing fully-clothed in the rain outside and have fish in my hand, putting it in my mouth enables me to eat it" and so on, the agent can learn to identify the world-states that correspond to the abstract concept of *having food available*, and then learn the appropriate action to take in *all* those states.
2. There are useful behaviors that need to be bootstrapped from lower-level concepts to higher-level ones in order to be learned. For example, newborns have an innate preference for looking at roughly face-shaped things ([Farroni et al. 2005](#)), which develops into a more consistent preference for looking at faces over the first year of life ([Frank, Vul & Johnson 2009](#)). One hypothesis is that this bias towards paying attention to the relatively-easy-to-encode-in-genes concept of "face-like things" helps direct attention towards learning valuable but much more complicated concepts, such as ones involved in a basic theory of mind ([Gopnik, Slaughter & Meltzoff 1994](#)) and the social skills involved with it.

Viewed in this light, *concepts are cognitive tools that are used for getting rewards*. At the most primitive level, we should expect a creature to develop concepts that abstract over situations that are similar with regards to the kind of reward that one can gain from taking a certain action in those states. Suppose that a certain action in state s_1 gives you a reward, and that there are also states $s_2 - s_5$ in which taking some specific action causes you to end up in s_1 . Then we should expect the creature to develop a common concept for being in the states $s_2 - s_5$, and we should expect that concept to be "more similar" to the concept of being in state s_1 than to the concept of being in some state that was many actions away.

"More similar" how?

In reinforcement learning theory, reward and value are two different concepts. The reward of a state is the actual reward that the reward function gives you when you're in that state or perform some action in that state. Meanwhile, the value of the state is the maximum total reward that you can expect to get from moving that state to others (times some discount factor). So a state A with reward 0 might have value 5 if you could move from it to state B, which had a reward of 5.

Below is a figure from DeepMind's recent *Nature* paper, which presented a deep reinforcement learner that was capable of achieving human-level performance or above on 29 of 49 Atari 2600 games ([Mnih et al. 2015](#)). The figure is a visualization of the representations that the learning agent has developed for different game-states in *Space Invaders*. The representations are color-coded depending on the value of the game-state that the representation corresponds to, with red indicating a higher value and blue a lower one.

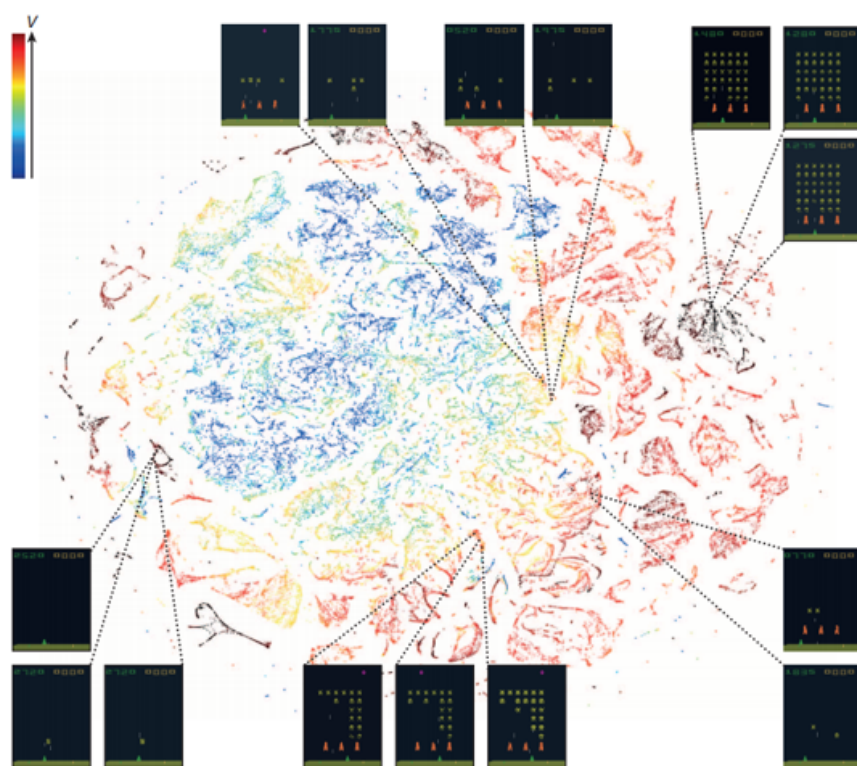


Figure 4 | Two-dimensional t-SNE embedding of the representations in the last hidden layer assigned by DQN to game states experienced while playing *Space Invaders*. The plot was generated by letting the DQN agent play for 2 h of real game time and running the t-SNE algorithm²⁵ on the last hidden layer representations assigned by DQN to each experienced game state. The points are coloured according to the state values (V , maximum expected reward of a state) predicted by DQN for the corresponding game states (ranging from dark red (highest V) to dark blue (lowest V)). The screenshots corresponding to a selected number of points are shown. The DQN agent

predicts high state values for both full (top right screenshots) and nearly complete screens (bottom left screenshots) because it has learned that completing a screen leads to a new screen full of enemy ships. Partially completed screens (bottom screenshots) are assigned lower state values because less immediate reward is available. The screens shown on the bottom right and top left and middle are less perceptually similar than the other examples but are still mapped to nearby representations and similar values because the orange bunkers do not carry great significance near the end of a level. With permission from Square Enix Limited.

As can be seen (and is noted in the caption), representations with similar values are mapped closer to each other in the representation space. Also, some game-states which are visually dissimilar to each other but have a similar value are mapped to nearby representations. Likewise, states that are visually similar but have a differing value are mapped away from each other. We could say that the Atari-playing agent has

learned a primitive concept space, where the relationships between the concepts (representing game-states) depend on their value and the ease of moving from one game-state to another.

In most artificial RL agents, reward and value are kept strictly separate. In humans (and mammals in general), this doesn't seem to work quite the same way. Rather, if there are things or behaviors which have once given us rewards, we tend to eventually start valuing them for their own sake. If you teach a child to be generous by praising them when they share their toys with others, you don't have to keep doing it all the way to your grave. Eventually they'll internalize the behavior, and start *wanting* to do it. One might say that the positive feedback actually modifies their reward function, so that they will start getting some amount of pleasure from generous behavior without needing to get external praise for it. In general, behaviors which are learned strongly enough don't need to be reinforced anymore ([Pryor 2006](#)).

Why does the human reward function change as well? Possibly because of the bootstrapping problem: there are things such as social status that are very complicated and hard to directly encode as "rewarding" in an infant mind, but which can be learned by associating them with rewards. One researcher I spoke with commented that he "wouldn't be at all surprised" if it turned out that sexual orientation was learned by men and women having slightly different smells, and sexual interest bootstrapping from an innate reward for being in the presence of the right kind of a smell, which the brain then associated with the features usually co-occurring with it. His point wasn't so much that he expected this to be the particular mechanism, but that he wouldn't find it particularly surprising if a core part of the mechanism was something that simple. Remember that incest avoidance [seems to bootstrap](#) from the simple cue of "don't be sexually interested in the people you grew up with".

This is, in essence, how I expect human values and human concepts to develop. We have some innate reward function which gives us various kinds of rewards for different kinds of things. Over time we develop a various *concepts* for the purpose of letting us maximize our rewards, and lived experiences also modify our reward function. Our *values* are concepts which abstract over situations in which we have previously obtained rewards, and which have become intrinsically rewarding as a result.

Getting back to conceptual expansion

Having defined these things, let's take another look at the two examples we discussed above. As a reminder, they were:

1. Children learn to represent dimensions such as "height" and "volume", as well as "big" and "bright", separately at around age 5.
2. As an inhabitant of the Earth, you've been used to people being unable to fly and landowners being able to forbid others from using their land. Then someone goes and invents an airplane, leaving open the question of the height to which the landowner's control extends.

I summarized my first attempt at describing the consequences of #1 as "it's a non-issue because none of the children's values have been defined in terms of the affected conceptual domain". We can now reframe it as "it's a non-issue because the [concepts that abstract over the world-states which give the child rewards] mostly do not make use of the dimension that's now been split into 'height' and 'volume'".

Admittedly, this new conceptual distinction might be relevant for estimating the value of a few things. A more accurate estimate of the volume of a glass leads to a more

accurate estimate of which glass of juice to prefer, for instance. With children, there probably is some intuitive physics module that figures out how to apply this new dimension for that purpose. Even if there wasn't, and it was unclear whether it was the "tall glass" or "high-volume glass" concept that needed be mapped closer to high-value glasses, this could be easily determined by simple experimentation.

As for the airplane example, I summarized my description of it by saying that "after an option became available which forced us to include a new dimension in our existing concept of landownership, we solved the issue by considering it in terms of our existing values". We can similarly reframe this as "after the feature of 'height' suddenly became relevant for the concept of landownership, when it hadn't been a relevant feature dimension for landownership before, we redefined landownership by considering which kind of redefinition would give us the largest amounts of rewarding things". "Rewarding things", here, shouldn't be understood only in terms of concrete physical rewards like money, but also anything else that people have ended up valuing, including abstract concepts like right to ownership.

Note also that different people, having different experiences, ended up making redefinitions. No doubt some landowners felt that the "being in total control of my land and everything above it" was a more important value than "the convenience of people who get to use airplanes"... unless, perhaps, they got to see first-hand the value of flying, in which case the new information could have repositioned the different concepts in their value-space.

As an aside, this also works as a possible partial explanation for e.g. someone being strongly against gay rights until their child comes out of the closet. Someone they care about suddenly benefiting from the concept of "gay rights", which previously had no positive value for them, may end up changing the value of that concept. In essence, they gain new information about the value of the world-states that the concept of "my nation having strong gay rights" abstracts over. (Of course, things don't always go this well, if their concept of homosexuality is too strongly negative to start with.)

The Flatland case follows a similar principle: the Flatlanders have some values that declared the inside of the rectangle a forbidden space. Maybe the inside of the rectangle contains monsters which tend to eat Flatlanders. Once they learn about 3D space, they can rethink about it in terms of their existing values.

Dealing with the AI in the box

This leaves us with the AI case. We have, via various examples, taught the AI to stay in the box, which was defined in terms of classical physics. In other words, the AI has obtained the concept of a box, and has come to associate staying in the box with some reward, or possibly leaving it with a lack of a reward.

Then the AI learns about quantum mechanics. It learns that in the QM formulation of the universe, "location" is not a fundamental or well-defined concept anymore - and in some theories, even the concept of "space" is no longer fundamental or well-defined. What happens?

Let's look at the human equivalent for this example: a physicist who learns about quantum mechanics. Do *they* start thinking that since location is no longer well-defined, they can now safely jump out of the window on the sixth floor?

Maybe some do. But I would wager that most don't. Why not?

The physicist cares about QM concepts to the extent that the said concepts are linked to things that the physicist values. Maybe the physicist finds it rewarding to develop a better understanding of QM, to gain social status by making important discoveries, and to pay their rent by understanding the concepts well enough to continue to do research. These are some of the things that the QM concepts are useful for. Likely the brain has some kind of causal model indicating that the QM concepts are relevant tools for achieving *those particular rewards*. At the same time, the physicist also has various other things they care about, like being healthy and hanging out with their friends. These are values that can be better furthered by modeling the world in terms of classical physics.

In some sense, the physicist knows that if they started thinking "location is ill-defined, so I can safely jump out of the window", then that would be changing the map, not the territory. It wouldn't help them get the rewards of being healthy and getting to hang out with friends - even if a hypothetical physicist who *did* make that redefinition would think otherwise. [It all adds up to normality.](#)

A part of this comes from the fact that the physicist's reward function remains defined over immediate sensory experiences, as well as values which are linked to those. Even if you convince yourself that the location of food is ill-defined and you thus don't need to eat, you will still suffer the negative reward of being hungry. The physicist knows that no matter how they change their definition of the world, that won't affect their actual sensory experience and the rewards they get from that.

So to prevent the AI from leaving the box by suitably redefining reality, we have to somehow find a way for the same reasoning to apply to it. I haven't worked out a rigorous definition for this, but it needs to somehow learn to care about being in the box *in classical terms*, and realize that no redefinition of "location" or "space" is going to alter what happens in the classical model. Also, its rewards need to be defined over models to a sufficient extent to avoid wireheading ([Hibbard 2011](#)), so that it will think that trying to leave the box by redefining things would count as self-delusion, and not accomplish the things it *really* cared about. This way, the AI's concept for "being in the box" should remain firmly linked to the classical interpretation of physics, not the QM interpretation of physics, because it's acting in terms of the classical model that has always given it the most reward.

It is my hope that this could also be made to extend to cases where the AI learns to think in terms of concepts that are totally dissimilar to ours. If it learns a new conceptual dimension, how should that affect its existing concepts? Well, it can figure out how to reclassify the existing concepts that are affected by that change, based on what kind of a classification ends up producing the most reward... when the reward function is *defined over the old model*.

Next post in series: [World-models as tools.](#)

Concept Safety: World-models as tools

The AI in the quantum box

In the [previous post](#), I discussed the example of an AI whose concept space and goals were defined in terms of classical physics, which then learned about quantum mechanics. Let's elaborate on that scenario a little more.

I wish to zoom in on a certain assumption that I've noticed in previous discussions of these kinds of examples. Although I couldn't track down an exact citation right now, I'm pretty confident that I've heard the QM scenario framed as something like "the AI previously thought in terms of classical mechanics, but then it finds out that the world *actually* runs on quantum mechanics". The key assumption being that quantum mechanics is in some sense *more real* than classical mechanics.

This kind of an assumption is a natural one to make if someone is operating on an AIXI-inspired model of AI. Although AIXI considers an infinite amount of world-models, there's a sense in which AIXI always strives to only have *one* world-model. It's always looking for the simplest possible Turing machine that would produce all of the observations that it has seen so far, while ignoring the computational cost of actually running that machine. AIXI, upon finding out about quantum mechanics, would attempt to update its world-model into one that only contained QM primitives and to derive all macro-scale events right from first principles.

No sane design for a real-world AI would try to do this. Instead, a real-world AI would take advantage of *scale separation*. This refers to the fact that physical systems can be modeled on a variety of different scales, and it is in many cases sufficient to model them in terms of concepts that are defined in terms of higher-scale phenomena. In practice, the AI would have a number of different world-models, each of them being applied in different situations and for different purposes.

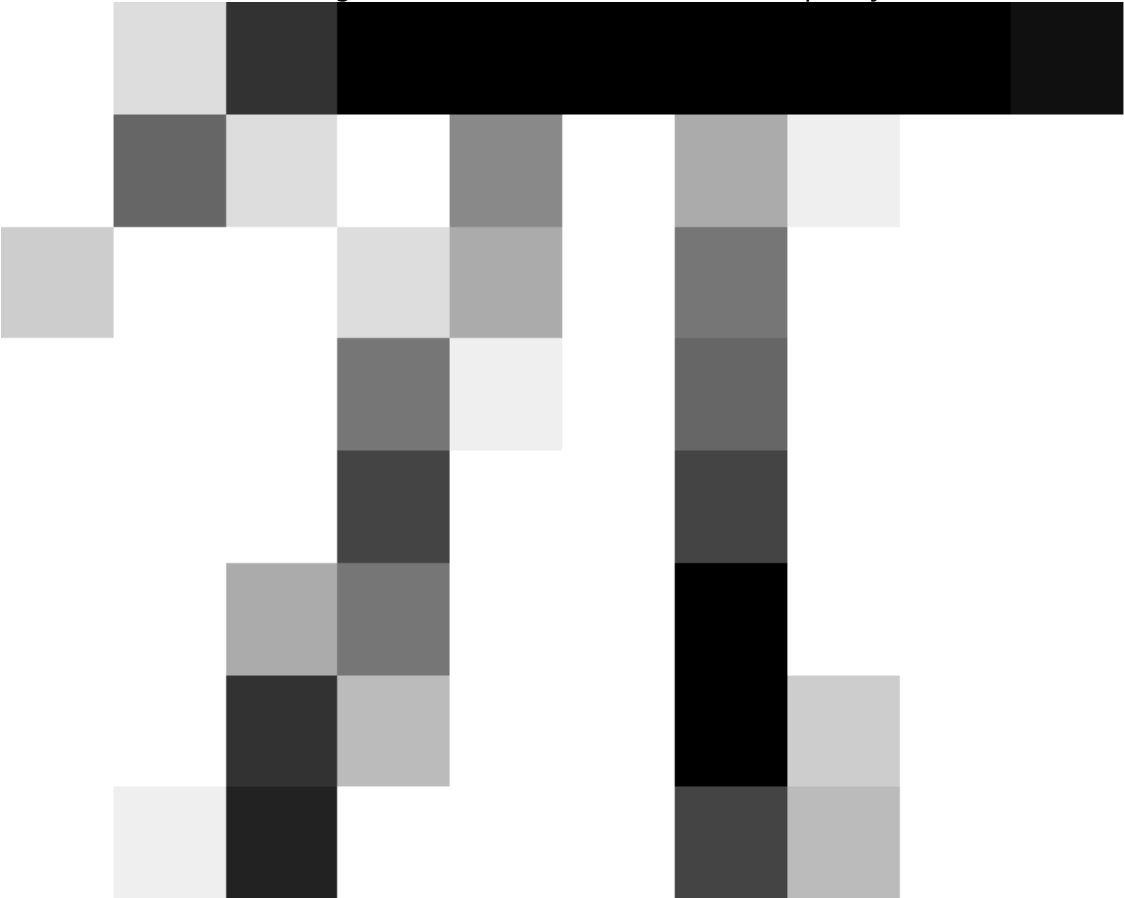
Here we get back to the view of concepts as tools, which I discussed in the previous post. An AI that was doing something akin to reinforcement learning would come to learn the kinds of world-models that gave it the highest rewards, and to selectively employ different world-models based on what was the best thing to do in each situation.

As a toy example, consider an AI that can choose to run a low-resolution or a high-resolution psychological model of someone it's interacting with, in order to predict their responses and please them. Say the low-resolution model takes a second to run and is 80% accurate; the high-resolution model takes five seconds to run and is 95% accurate. Which model will be chosen as the one to be used will depend on the cost matrix of making a correct prediction, making a false prediction, and the consequence of making the other person wait for an extra four seconds before the AI's each reply.

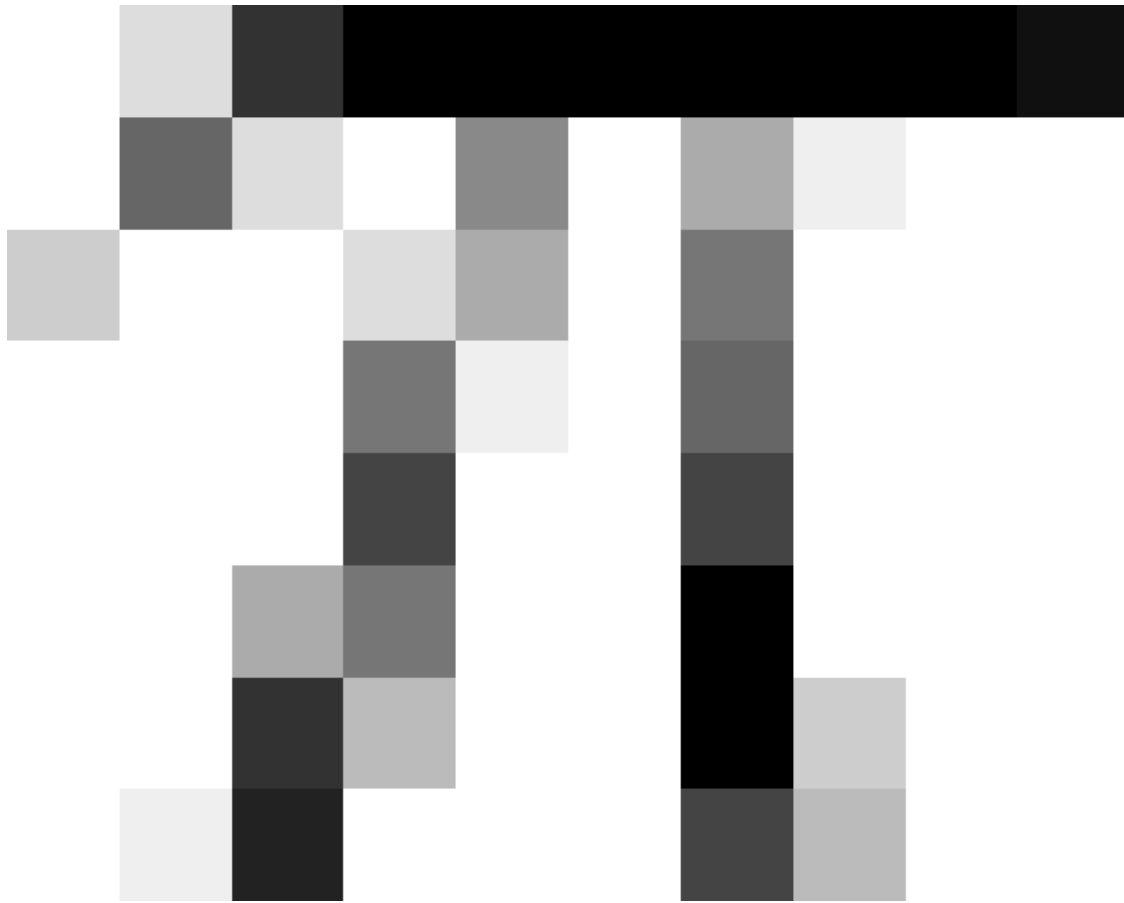
We can now see that a world-model being the most real, i.e. making the most accurate predictions, doesn't automatically mean that it will be used. It also needs to be fast enough to run, and the predictions need to be useful for achieving something that the AI cares about.

World-models as tools

From this point of view, world-models are *literally* tools just like any other. [Traditionally](#) in reinforcement learning, we would define the value of a policy



in state s as the expected reward given the state s and the policy




$$V^{\pi}(s) = E[R|s, \pi]$$

but under the "world-models are tools" perspective, we need to also condition on the world-model m ,

$$V^{\pi}(s, m) = E[R|s, \pi, m]$$

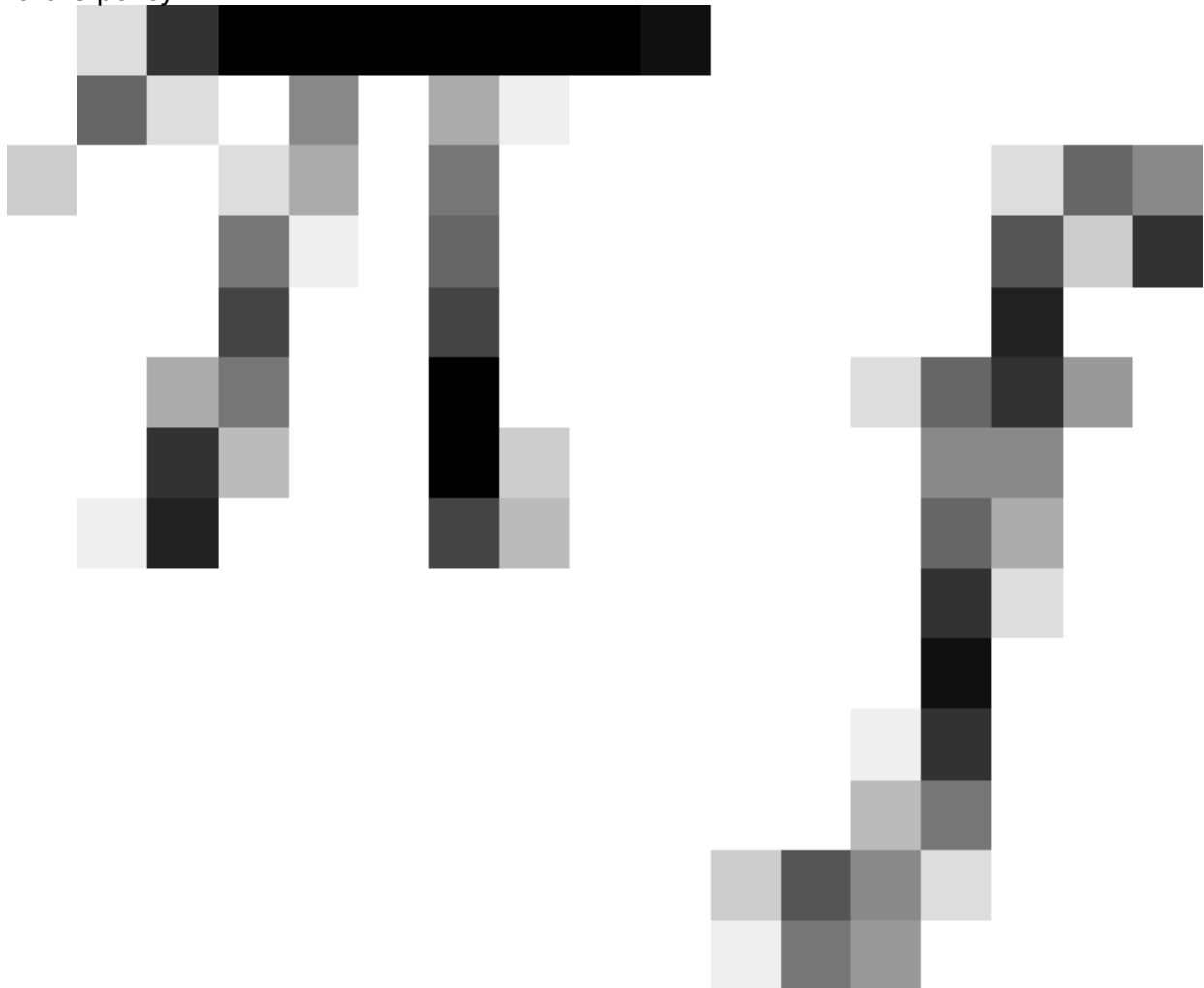
We are conditioning on the world-model in several distinct ways.

First, there is the expected *behavior of the world as predicted by world-model m* . A world-model over the laws of social interaction would do poorly at predicting the movement of celestial objects, if it could be applied to them at all. Different predictions of behavior may also lead to differing predictions of the *value* of a state. This is described by the equation above.

Second, there is the expected *cost of using the world-model*. Using a more detailed world-model may be more computationally expensive, for instance. One way of interpreting this in a classical RL framework would be that using a specific world-model will place the agent in a different state than using some other world-model. We might describe by saying that in addition to the agent choosing its next action a on each time-step, the agent also needs to choose the world-model m which it will use to analyze its next observations. This will be one of the inputs for the transition function  to the next state.

$$\delta : s \times (a, m) \rightarrow (s, m)$$

Third, there is the expected *behavior of the agent using world-model m* . An agent with different beliefs about the world will act differently in the future: this means that the future policy



actually depends on the chosen world-model.

$$P(\pi_f) = P(\pi|m)$$

Some very interesting questions pop up at this point. Your currently selected world-model is what you use to evaluate your best choices for the next step... including the choice of what world-model to use next. So whether or not you're going to switch to a different world-model for evaluating the next step depends on whether your current world-model says that a different world-model would be better in that step.

We have not fully defined what exactly we mean by "world-models" here. Previously I gave the example of a world-model over the laws of social interaction, versus a world-model over the laws of physics. But a world-model over the laws of social interaction, say, would not have an answer to the question of which world-model to use for things it couldn't predict. So one approach would be to say that we actually have some meta-model over world-models, telling us which is the best to use in what situation.

On the other hand, it does also seem like humans often use a specific world-model and its predictions to determine whether to choose another world-model. For example, in rationalist circles you often see arguments to the line of, "self-deception might give you extra confidence, but it introduces errors into your world-model, and in the long term those are going to be more harmful than the extra confidence is beneficial". Here you see an implicit appeal to a world-model which predicts an accumulation of false beliefs with some specific effects, as well as predicting the extra self-esteem with its effects. But this kind of an analysis incorporates very specific causal claims from various (e.g. psychological) models, which are models over the world rather than just being part of some general meta-model over models. Notice also that the example analysis takes into account the way that having a specific world-model affects the state transition function: it assumes that a self-deceptive model may land us in a state where we have a higher self-esteem.

It's possible to get stuck in one world-model: for example, a strongly non-reductionist model evaluating the claims of a highly reductionist one might think it obviously crazy, and vice versa. So it seems that we do need *something* like a meta-evaluation function. Otherwise it would be too easy to get stuck in one model which claimed that it was the best one in every possible situation, and never agreed to "give up control" in favor of another one.

One possibility for such a thing would be a relatively model-free learning mechanism, which just kept track of the rewards accumulated when using a particular model in a particular situation. It would then bias the selection of the model towards the direction of the model that had been the most successful so far.

Human neuroscience and meta-models

We *might* be able to identify something like this in humans, though this is currently very speculative on my part. Action selection is [carried out in the basal ganglia](#): different brain systems send the basal ganglia "bids" for various actions. The basal ganglia then chooses which actions to inhibit or disinhibit (by default, everything is inhibited). The basal ganglia also implements reinforcement learning, selectively strengthening or weakening the connections associated with a particular bid and context when a chosen action leads to a higher or lower reward than was expected. It

seems that in addition to choosing between motor actions, the basal ganglia also chooses between different cognitive behaviors, likely even *thoughts*:

If action selection and reinforcement learning are normal functions of the basal ganglia, it should be possible to interpret many of the human basal ganglia-related disorders in terms of selection malfunctions. For example, the akinesia of Parkinson's disease may be seen as a failure to inhibit tonic inhibitory output signals on any of the sensorimotor channels. Aspects of schizophrenia, attention deficit disorder and Tourette's syndrome could reflect different forms of failure to maintain sufficient inhibitory output activity in non-selected channels. Consequently, insufficiently inhibited signals in non-selected target structures could interfere with the output of selected targets (expressed as motor/verbal tics) and/or make the selection system vulnerable to interruption from distracting stimuli (schizophrenia, attention deficit disorder). The opposite situation would be where the selection of one functional channel is abnormally dominant thereby making it difficult for competing events to interrupt or cause a behavioural or attentional switch. Such circumstances could underlie addictive compulsions or obsessive compulsive disorder. ([Redgrave 2007](#))

Although I haven't seen a paper presenting evidence for this particular claim, it seems plausible to assume that humans similarly come to employ new kinds of world-models based on the extent to which using a particular world-model in a particular situation gives them rewards. When a person is in a situation where they might think in terms of several different world-models, there will be neural bids associated with mental activities that recruit the different models. Over time, the bids associated with the most successful models will become increasingly favored. This is also compatible with what we know about e.g. [happy death spirals](#) and [motivated stopping](#): people will tend to have the kinds of thoughts which are rewarding to them.

The physicist and the AI

In my previous post, when discussing the example of the physicist who doesn't jump out of the window when they learn about QM and find out that "location" is ill-defined:

The physicist cares about QM concepts to the extent that the said concepts are linked to things that the physicist values. Maybe the physicist finds it rewarding to develop a better understanding of QM, to gain social status by making important discoveries, and to pay their rent by understanding the concepts well enough to continue to do research. These are some of the things that the QM concepts are useful for. Likely the brain has some kind of causal model indicating that the QM concepts are relevant tools for achieving those particular rewards. At the same time, the physicist also has various other things they care about, like being healthy and hanging out with their friends. These are values that can be better furthered by modeling the world in terms of classical physics. [...]

A part of this comes from the fact that the physicist's reward function remains defined over immediate sensory experiences, as well as values which are linked to those. Even if you convince yourself that the location of food is ill-defined and you thus don't need to eat, you will still suffer the negative reward of being hungry. The physicist knows that no matter how they change their definition of the world, that won't affect their actual sensory experience and the rewards they get from that.

So to prevent the AI from leaving the box by suitably redefining reality, we have to somehow find a way for the same reasoning to apply to it. I haven't worked out a rigorous definition for this, but it needs to somehow learn to care about being in the box in classical terms, and realize that no redefinition of "location" or "space" is going

to alter what happens in the classical model. Also, its rewards need to be defined over models to a sufficient extent to avoid wireheading (Hibbard 2011), so that it will think that trying to leave the box by redefining things would count as self-delusion, and not accomplish the things it really cared about. This way, the AI's concept for "being in the box" should remain firmly linked to the classical interpretation of physics, not the QM interpretation of physics, because it's acting in terms of the classical model that has always given it the most reward.

There are several parts to this.

1. The "physicist's reward function remains defined over immediate sensory experiences". Them falling down and breaking their leg is still going to hurt, and they know that this won't be changed no matter how they try to redefine reality.
2. The physicist's *value* function also remains defined over immediate sensory experiences. They know that jumping out of a window and ending up with all the bones in their body being broken is going to be really inconvenient even if you disregarded the physical pain. They still cannot do the things they would like to do, and they have learned that being in such a state is non-desirable. Again, this won't be affected by how they try to define reality.

We now have a somewhat better understanding of what exactly this means. The physicist has spent their entire life living in the classical world, and obtained nearly all of their rewards by thinking in terms of the classical world. As a result, using the classical model for reasoning about life has become strongly selected for. Also, the physicist's classical world-model predicts that thinking in terms of that model is a very good thing for surviving, and that trying to switch to a QM model where location was ill-defined would be a very bad thing for the goal of surviving. On the other hand, thinking in terms of exotic world-models remains a rewarding thing for goals such as obtaining social status or making interesting discoveries, so the QM model does get more strongly reinforced in that context and for that purpose.

Getting back to the question of how to make the AI stay in the box, ideally we could mimic this process, so that the AI would initially come to care about staying in the box. Then when it learns about QM, it understands that thinking in QM terms is useful for some goals, but if it were to make itself think in purely QM terms, that would cause it to leave the box. Because it is thinking mostly in terms of a classical model, which says that leaving the box would be bad (analogous to the physicist thinking mostly in terms of the classical model which says that jumping out of the window would be bad), it wants to make sure that it will continue to think in terms of the classical model when it's reasoning about its location.