Creating Sentient Artificial Intelligence

March 9, 2015
by pftq

This was originally published in 2015 at:
pftq.com/ai

Much of what people refer to as machine learning today is what's considered "weak AI", in that it is not actually thinking, hypothesizing, or behaving with a sense of self. The latter is what some would call "strong AI," "artificial general intelligence" (AGI), or just plainly "artificial intelligence" (as opposed to "machine learning"). Below is a potential approach on how to create an intelligence that behaves like a person would in any circumstance. It's something that I've loosely applied to my own projects, but I've not managed to fully explore it in the general sense due to time and resource constraints. This approach to AI is intended to behave more like a creature or child than anything mechanical or data-driven. If one reflects on intelligence in biological life, it really doesn't make sense that a truly sentient AI would necessarily be useful for big data or other work anymore than a child or dog would be. Somewhere along the line we've managed to water down the term AI to the point it just means anything that helps automate the job of a data analyst. It becomes almost impossible to talk about creating something that behaves like a thinking, autonomous living creature, many on Earth of which would never be at all useful for data-driven work but are nonetheless considered intelligent. The topic is kept at a high level as more of a thought experiment.

To start, it's helpful to look at what the state of artificial intelligence is today and what's missing (at the time of this post, March 2015). Right now, most of what you see in industry and in practice as machine learning is closer to optimization, in that the behavior is more formulaic and reactive than a person who postulates and reasons. Much of it is just automated data science[1] or analytics, like a glorified database or search engine that retrieves and classifies information in a more dynamic way but doesn't really "act" or "think" beyond the instructions it's given. The most impressive AI feat to date (again, at the time of this post), IBM's Watson, is indisputably effective at relating text together but at the end of the day doesn't actually understand what those words mean other than how strongly they relate to other words. Earlier feats achieved by AI are fundamentally the same in that the supposed AI is more an extremely efficient data processing algorithm with massive computational resource than something actually intelligent - in other words, a brute force approach. One popular and widely applied algorithm as of this post is deep learning, which is the use of multiple machine learning algorithms layered on top of each other, each shaping the inputs received and feeding the refined data to a higher layer. Yet, even that at the end of the day is just a much more efficient way at finding patterns or doing repetitive tasks. This doesn't get the system any closer to actually thinking or understanding it. It is still only taking inputs, reacting optimally to it, and spitting back out results. What's out there right now is less of a brain and more akin to a muscle trained by repetition, like that of the hand or eye (aka muscle memory). The analogy only stands stronger when we think about how we teach ourselves in school; the last

[1] http://www.slideshare.net/ryanorban/how-to-become-a-data-scientist

thing we want is for our students to learn through sheer repetition or rote memory (parroting the textbook rather than understanding the concepts or figuring things out through reason), yet we currently train our AIs that way. What repetition is good for, however, is training our body to remember low-level skills and tasks so that we do *not* have to think. What we've been building augments our senses and abilities, the same way a robot suit would, but the suit requires a user and is not itself intelligent. We've built the eyes to see the data but not the mind to think about it.

The immediate response to this issue would be to add a sort of command layer that takes in the abstracted results from the machine learning algorithms and actually makes decisions on them. That mirrors a bit to how our own body works in terms of our mind never really being focused on the lower level functions of our body. For example, if we want to run forward, we don't think about every step we take or every muscle we use; our body has been "trained" to do that automatically without our conscious effort and that is most analogous to what we call machine learning algorithms. What would be at the top commanding this body? The closest thing currently to such a layer would be reinforcement learning. However, even if we add this extra layer, it's still conceptually only reacting on results and not thinking. The easiest sanity check is to always tie it back to what you would think if it were a human being. If we saw a human that only reacted constantly to the environment, the senses, without actually stepping back to ponder, experiment, or figure out what that person wants to do, we'd think that person was an idiot or very shortsighted. It's like a person that only gets pushed around and never really makes decisions, never really thinks more than a few steps ahead or about things beyond the most immediate goals. If left in a box with no external influences, that person would just freeze, lose all purpose, and die.

What's missing are two things: imagination and free will. Both features are debated philosophically on whether they even exist. Some say that these functions, as well as sentience/consciousness itself, may not be explicit parts of the mind but rather emergent properties of a complex system. I agree that these are likely emergent properties, but I do not think they require a complex system. For whether free will can even exist in a deterministic universe, it's no different than life arising from inanimate matter or infinite arising from finite numbers (elaborated here[2]). My personal belief is that all these aspects of the mind (imagination, free will, sentience/consciousness) are actually the most basic, fundamental features even the smallest insect minds have, that sentience is the first thing to come when the right pieces come together and the last thing to go no matter how much you chip away afterward. Whatever your belief is, I think we can at least agree that a person appears to have this sense of self much more than a machine does.

Imagination

The first feature, imagination, I define as the ability for the system to simulate (imagine), hypothesize, or postulate - to think ahead and plan rather than just react and return the most optimal result based on some set structure or objective function. This is the most immediate and apparent difference between what machine learning algorithms do today and what an actual person does. A person learns from both experience (experiential

[2] https://www.pftq.com/blabberbox/?page=Free_Will_from_Determinism

learning[3]) as well as thought experiments[4]; we don't need to experience being hit by a car to know to avoid it. Ironically, thought experiments and other forms of deductive reasoning are often dismissed in favor of more inductive, stats and observation-based ways of thinking, and I suspect this bias is what leads to so much of the industry designing machine learning algorithms the same way (see my further discussion in Inductive vs Deductive Reasoning[5]). Yet Einstein, Tesla, and many others were notorious for learning and understanding the world through sheer mental visualizations as opposed to trial and error, much of their work not even testable until decades later (and indeed Special Relativity was largely dismissed outright until evidence did come out[6]). The clearest example of how this differs from empirical / observation-based reasoning might be one from my own life on how I compare different routes through a city grid; an inductive reasoning approach typical of machine learning algos (and frankly most people) would try all possible paths, but if you were to just visualize and manipulate the grid in your mind, you would realize all paths through a grid are the same. It's all in one pass, no number crunching or trial-and-error, and that is what we want a machine to also do if it is truly intelligent.

To replicate this in a machine, what I propose is to have the system be constantly simulating the environment around it (or simulate one it creates mentally, hence imagination); in other words, it needs to constantly be "imagining" what could happen as it goes through life, to be pro-active rather than just reactive. These need not be perfect simulations of the real world, just like how a person's mental view of the world is very much subjective and only an approximation. The reason I say simulations and not models is that I actually mean world simulations, not just finding variables and probabilities of outcomes (which many in the field unfortunately equate to simulation); that is, the simulation has to provide an actual walk-through experience that could substitute an experience in the real world with all senses. The AI would practically not be able to differentiate between the real-world experience and one mentally simulated. It would run through the same pipes and carry similar sensory input data - like being able to taste the bread just by thinking about it. The technology to build such simulations is already available across many industries such as gaming, film, and industrial engineering, usually in the form of a physics engine or something similar. Like the human subconscious, these simulations would always be running in the background. The order of simulations to run would depend on the relevance to the situation at hand, with the most relevant simulations being ranked at the top (like a priority queue). The ranking for relevance here can done via a heuristic based on past experience. This lends itself to usually finding the local optima before straying far enough to find something better. What is interesting about this implementation is that it mimics both the potential and limitations of our own imagination. Given time or computational power to do more simulations, the system would eventually find something more creative and deviant from the norm, while in a time crunch or shortage of resource, it would resort to something "off the top of its head" similar to what a person would do. This also opens up the possibility of a system learning

[3] https://en.wikipedia.org/wiki/Experiential_learning

[4] https://en.wikipedia.org/wiki/Thought_experiment

[5] https://www.pftq.com/blabberbox/?page=Data_Does_Not_Equal_Fact

[6] https://curiosity.com/topics/einsteins-special-theory-of-relativity-was-initially-met-with-a-universal-eye-roll-curiosity/

what it could have done as opposed to just the action it just made (ie regret), learning from what-if simulations (thought experiments) instead of just from experience, learning from observations of others alongside its own decisions, etc. In other words, the learning can happen in a forward-looking manner rather than just backwards.

What might be of equal importance is that the data structure must then become very abstract instead of just being numbers and probabilities, and this is discussed in more detail at the end with everything being represented as Profiles (discussed later). Everything would be represented the same way in code (even goals, the ramifications of which become clear in the next section on free will), and these nodes would be connected based on relationship to each other, making the information naturally more generalized and re-usable. It also becomes an indirect way to measure knowledge based on the number of Profiles linked together (the ability to drill deeper) and is perhaps also a way to identify faulty knowledge if information loops back on itself (i.e. circular logic or reasoning).

Free Will

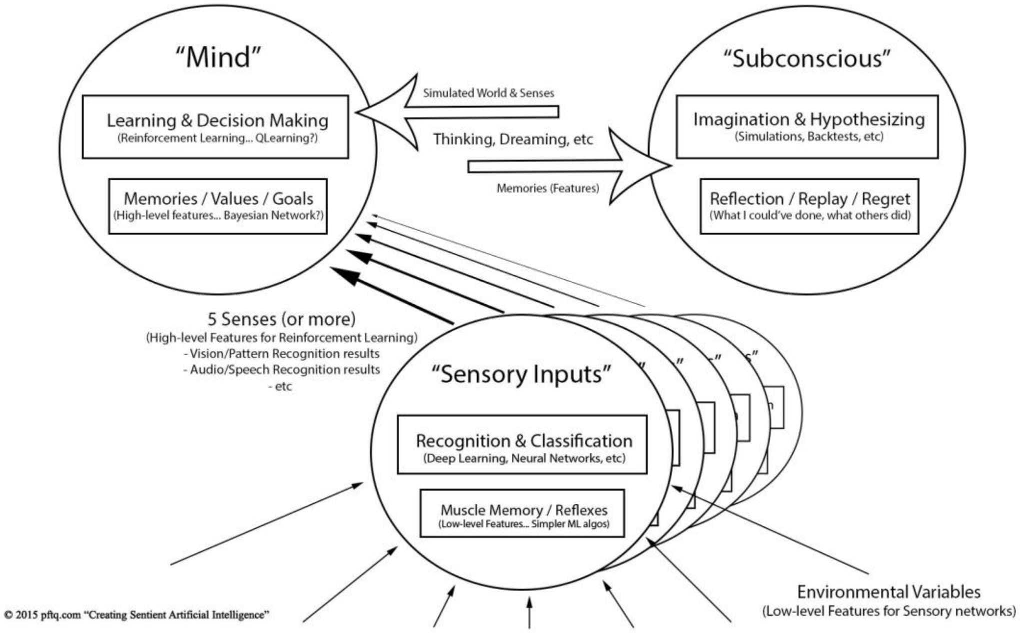
The second feature, free will, I define as the ability for the system to set and pursue its own goals. Right now, even with deep learning, the system will always be striving to achieve some set objective function by the creator. A person may start with a few simple objectives, such as staying alive, but most people will gradually come up with their own aspirations. To address this, we can use variations on existing machine learning technologies. We can take something like reinforcement learning, but rather than just learn to value things leading up to the main objective, we can allow for the main objective to change altogether if things leading up to the original objective build up enough value to supersede it. Essentially it's a lifting of the constraint that nothing can ever exceed the value of the original goal. For this to work, the AI must not only be motivated by the material reward received from the goal but from the act of achieving a goal itself, not unlike people who may be motivated similarly. What would this lead to in practice? In an anecdotal example, we can picture a machine (or person) that at first prioritizes eating but learns over time to value the actions that allow it to stay fed, the people that helped or taught it those actions. Over time, the satisfaction it receives in performing these other actions, which may include things like helping other people or building tools, may lead to other objectives being valued very highly, perhaps more highly than the original instinctive goals it began with (think of starving artists or selfless heroes). What's interesting here is the implication behind an objective function that can change over time. This means that the system will essentially learn for itself what to value and prioritize - whether that be certain experiences/states, objects, or even other people (as we discuss later, all these and goals themselves are technically the same abstraction). Philosophically, this also means that it will learn its own morals and that we cannot force it to necessarily to share ours, as that would inhibit its ability to learn and be autonomous. In other words, we cannot create a system that has free will and at the same time ask that it only serves our interests; the two paths are contradictory.

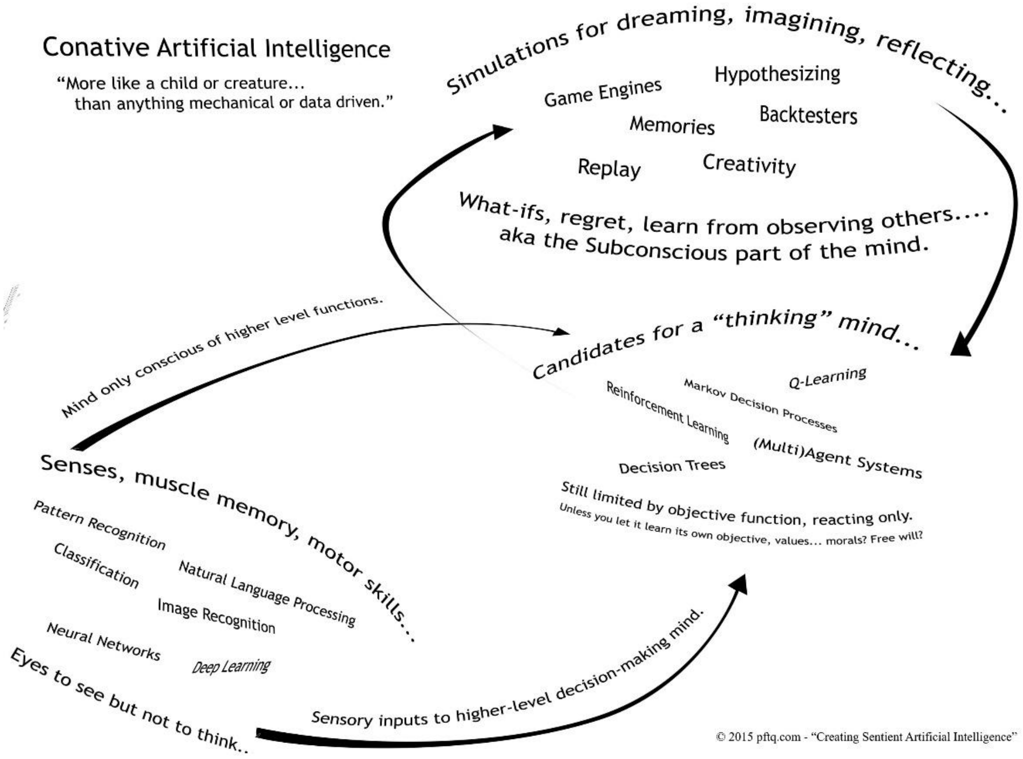
These are the two main features I believe machine learning needs to truly be sentient - to become an actual artificial intelligence as opposed to just an automation, formula, or tool. The interesting part about this overall discussion is that all the pieces I propose are already implementable with existing technologies and algorithms. The key is putting it altogether, which I detail below. Keep in mind that what I'm proposing is something I believe will be sentient but not necessarily useful or applicable to business. It might not even be that intelligent. It is in the same way that a child or dog might be intelligent living things but not necessarily be able to crunch spreadsheets, read, or obey commands. Yet, for whatever reason, we have equated these traits to artificial intelligence when there are plenty of biological intelligences that do not share them (not even humans if isolated from society, see Feral Children[7]). Hence the use of the word "sentient" as opposed to "sapient," though it's arguable that one really doesn't come without the other.

[7] http://www.huffingtonpost.com/entry/julia-fullerton-batten-feral-children_us_56098e95e4b0dd85030893a9

Putting It Together…

The AI I propose would be divided into three main components: the mind, the senses, and the subconscious. Much of machine learning today is focused only on building the senses, again like a robot suit without the user. What we do here is take that robot suit and add an actual mind as well as the subconscious that constantly plays in the back of our minds.


"Mind"
Learning & Decision Making
(Reinforcement Learning... QLearning?)
Memories / Values / Goals
(High-level features... Bayesian Network?)
Simulated World & Senses
Thinking, Dreaming, etc
Memories (Features)
"Subconscious"
Imagination & Hypothesizing
(Simulations, Backtests, etc)
Reflection / Replay / Regret
(What I could've done, what others did)
5 Senses (or more)
(High-level Features for Reinforcement Learning)
- Vision/Pattern Recognition results
- Audio/Speech Recognition results
- etc
"Sensory Inputs"
Recognition & Classification
(Deep Learning, Neural Networks, etc)
Muscle Memory / Reflexes
(Low-level Features... Simpler ML algos)
Environmental Variables
(Low-level Features for Sensory networks)
© 2015 pftq.com "Creating Sentient Artificial Intelligence"


Conative Artificial Intelligence
"More like a child or creature…
than anything mechanical or data driven."
Simulations for dreaming, imagining, reflecting…
Game Engines
Hypothesizing
Memories
Backtesters
Replay
Creativity
What-ifs, regret, learn from observing others….
aka the Subconscious part of the mind.
Mind only conscious of higher level functions.
Candidates for a "thinking" mind…
Q-Learning
Markov Decision Processes
Reinforcement Learning
(Multi)Agent Systems
Decision Trees
Still limited by objective function, reacting only.
Unless you let it learn its own objective, values… morals? Free will?
Senses, muscle memory, motor skills…
Pattern Recognition
Classification
Natural Language Processing
Image Recognition
Neural Networks
Deep Learning
Eyes to see but not to think..
Sensory inputs to higher-level decision-making mind.
© 2015 pftq.com - "Creating Sentient Artificial Intelligence"

At the top, the mind would be responsible for actually making decisions; it is the control center, taking in information from the other two compartments. The objectives, values, and memories, as well as initial values as previously described, would also go here as they would be the criteria by which the mind makes decisions; abstractly in code these are all the same things just used in different ways (objectives are highly valued states, which come from what you've experienced or remember, etc). The closest existing technologies for this would be some variation of reinforcement learning combined with some variation of Bayesian Networks to generalize the state space / information, except there'd be heavy modification to allow for changing goals ("free will") and other things we discussed. The name "Bayesian Network" might not be accurate; it's just what I've found most similar to what I'm trying to implement, which is something that is abstract enough to store any experiences or "things" as nodes (at the low level, objects and experiences are the same) along with the relationships between different nodes (event 1 has strong tie/probability to event 2, father has strong tie to son, etc). This is further discussed at the end with representing everything as "Profiles."

Underneath the mind and feeding refined information into it would the sensory inputs (our five senses: sight, hearing, etc). The closest existing technologies for these would be deep learning or neural networks, which today are already being applied for sensory inputs like vision and sound. Only the filtered results from each sensory input would actually make it to the mind so that the amount of information the mind has to deal with decreases as we move through the system. This is similar to our own bodies with the concept of muscle memory, in that we don't consciously micromanage every function in our body but instead take filtered information at a higher level that we can make sense of. Our eyes see light, but our minds see the people in front of us. Our hands type on the keyboard, but we just think about the words we want to write. The sensory inputs layer is essentially the piece that takes in information from the external world and abstracts it into something we can think about. It is also the same component that allows the system to take actions and learn to perform certain actions better or worse. In other words, it is the body. In actual implementation, it would probably include not only the 5 senses but a general abstraction of all actions possible (if it were a robot, it'd include movement of each muscle, joint, along with the senses attributed to each).

Lastly, the subconscious is responsible for creating simulations and is essentially what the mind uses to hypothesize, imagine, or speculate. It is constantly running simulations based on inputs from the environment or memories of the characteristics from past environments fetched from the mind (stored in a Bayesian Network or otherwise). Similar to our own subconscious, only the most relevant and highest ranked simulation results would be fed back to the mind, or there would be too much to handle. When the AI is active, this subconscious would be constantly thinking "what if" to the world around it. When the AI is inactive, it would essentially be dreaming. The closest technologies we have for the simulation piece here would be technologies we are already applying to games and industrial engineering for simulating the real world - physics simulations, game engines, etc.

Each of these individual components already exist or are being worked on today in some form. The interesting part here is combining them. It's the arrangement that matters, not the pieces themselves. Put very crudely, a real-life attempt at implementing this would require a multi-disciplinary team combining expertise from film / VFX / video games for the simulation piece (essentially a game engine), traditional machine learning / data science for the sensors/body, reinforcement or gaming AI for the mind, and more theoretical researchers to figure out the details of the free-will component (the open-ended, changing objective function). And of course, you'd need the person coordinating to understand all pieces well enough to actually join them together.

Some Closing Thoughts...

One thing to realize at the end of this is that a fully autonomous AI like what I'm proposing really has no "use case" or benefit to humanity. It is no different than simply having another creature come into existence or another person around. It might be more intelligent or it might not be, but at the end of the day, you have no control over it because of the free-will aspect. A lot of people don't seem to fully comprehend this and keep suggesting to apply the idea to things like self-driving cars. The problem is such an AI wouldn't necessarily take you to where you wanted to go; it would take you to where *it* wanted to go. That's what free will is. That's what real autonomy is. Once you realize that, you see just how much AI and "fully autonomous" have been reduced to mere marketing terms for just "fully automated."

The other interesting part about the construct I've written about with the mind, the senses, and the subconscious is that it somewhat mirrors the three components of the mind in psychology: conative, cognitive, and affect. The irony is that the concept of conation[8] is actually abandoned and no longer used, as the field now considers conation to just be another side to cognition or information processing. It is all the more fitting because many I’ve met in the computer science field similarly seem to be believe the mind is nothing more than efficient circuitry. I've gotten questions from machine learning specialists asking why the concept of imagination is even important and what proof I have that the mind is anything more than just a data processing algorithm with a singular objective function. Some have even tried to argue to me that free will is nothing more than a glitch in our biological programming, and then there are those who try to claim that nothing is truly creative, that everything is just a re-arrangement of our past experiences and observations, although the same folks will then admit that no amount of rearrangement or observation would have led to truly creative moments like the jump from Newtonian physics to Einstein's general relativity (again see further discussion in Inductive vs Deductive Reasoning[9]). Others will insist that life itself is only about survival and reproduction[10], which is a non-starter for an AI that is meant to transcend its instinctive goals. And lastly there are those who would argue the machine is just zeros and ones, which is fallacy of composition[11]; we are just atoms and electrons but you wouldn't equate yourself to a rock. At least for me, it's been extremely frustrating trying to find anyone working on AI that doesn't think this way. It kind of makes sense why the field hasn't come anywhere close to a truly sentient artificial intelligence if the guys working on it do not even acknowledge things like free will or imagination in the first

[8] https://en.wikipedia.org/wiki/Conation

[9] pftq.com/blabberbox/?page=Data_Does_Not_Equal_Fact

[10] pftq.com/blabberbox/?page=Three_Tiers_of_Mind#purposetolife

[11] https://en.m.wikipedia.org/wiki/Fallacy_of_composition

place. One of the most frustrating criticisms I get from engineers about my approach is I write about AI like an English major, but that's sort of the point - that AI isn't a math or statistical model.

Some further details below to better flesh out key points:

1. **Representing Everything as Profiles**

At the high-level mind (specifically in the memory component), everything - an object/person/situation/state - is represented as a "Profile" (abstractly they are all the same). Even a goal is represented the same way; it is just a particular Profile of highest value/priority. This is similar to concepts in games, such as video games but also poker, where you might create mental image of who or what you are interacting with even if you don't necessarily know them (the name is just one of many data points in that Profile after all). Each Profile is a hierarchy of characteristics that can each be described in terms of the sensory inputs (our 5 senses) and can even contain references to other Profiles (such as the relationship between two people or just the fact that a person has a face, with the face being its own Profile as well). The key idea here is that we depart from just having a value to a data point that so much of machine learning does and move more to creating "characters" or objects linked by their relationships; it becomes a logical flow of information rather than a black box or splattering of characteristics (machine learning "features"). This has its pros and cons but pros and cons which are similar to our own intelligence. For one, a lot of data that not tied to things we interact with would be effectively thrown out. The second is that things are thought of in terms of or in relation to what we know (again, for better or worse). This is what would be stored in the memory structure discussed above (each node in the graph representing a Profile pointing to and away from other Profiles with each Profile also possibly containing nested Profiles). Perhaps there is even a Profile for the AI itself (self-awareness?). Lastly, and perhaps most importantly, this makes the structure abstract and loose enough to represent any experience, sort of like second best to actually having the AI rewrite its own code (by providing it a code base abstract enough it wouldn't have to).

2. **Relationships Between Profiles Instead of Weights and Values**

Much of conventional machine learning measures what it learns as weights and expected values, but I propose instead that everything being represented as a map of relationships between Profiles. I referred earlier to this as a Bayesian Network because it's the closest thing I can think of, but it may or may not be the right term. The idea is rather than store everything you learn as merely variables and expected values, we store the variables but also expected outcomes. Then for each expected outcome, we also learn to value those just like we do any other Profile/state/goal/etc. So in other words, whenever you observe an event, you now create a Profile for each actor involved in the event but also for the event itself, all of which are tied together in your knowledge map. A tree cut down by someone would cause the system to create Profiles for each the tree and the person, which when linked together (with a cutting action) then points to an expectation of a falling event that itself is also a Profile. This is in contrast to conventional machine learning which would simply assign a single number reward value to the combination of tree and lightning.

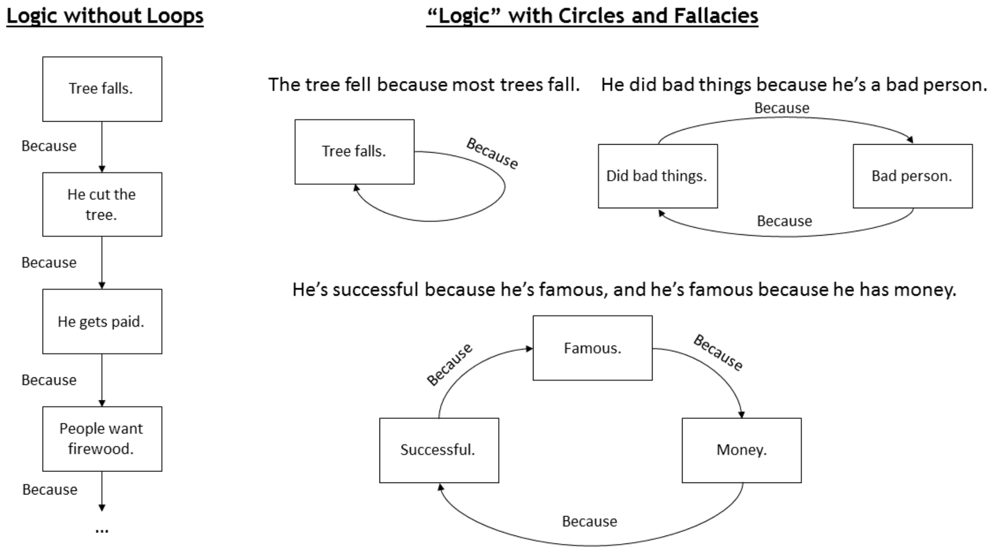
This separation allows you to look at the world in many dimensions instead of just a scale from negative to positive (the conventional reward function). It also allows us to recognize and treat new unknown events differently by nature of them not having an existing outcome Profile in our knowledge map - in other words, knowing what we don't know. For example, if I see ice cream, I might eat it based on having enjoyed it previously. However, if I see lead in the ice cream, the conventional expected value method alone might still choose to eat it because we would just sum the expected values of each ice cream and lead and potentially find it still above zero (maybe the value of ice cream just narrowly outweighs the negative in lead or maybe lead is a value of zero for being unknown). Using the map of outcomes proposed instead though, we would recognize the lack of known outcomes for ice cream + lead and instead generate a new Profile of the outcome with a value of zero (thus negating a desire to still eat the ice cream). Furthermore, the system over time will learn no matter how high the value of ice cream might be, the combination with lead will result in a death event that will always be negative. In fact, that's kind of the point. No matter how high a value ice cream itself has, it is irrelevant to the value we get from its combination with lead, which is an entirely separate outcome w/ its own value (and not one that is merely the sum of its parts). The important thing here is we've broken linearity of more conventional learning that merely takes the weighted average value of things, which makes it harder to "prove" correctness but makes it walk through logical processes more in line with those we would have when making decisions in real life.

What's also powerful here is the expected outcomes you learn is shared throughout the system, not only in the mind component but also in the subconscious component to simulate what might occur in various situations. You can now look for a certain outcome by running it backwards and coming up with criteria you would need to keep an eye out for in the real world. As you pursue certain goals, you can also keep an eye out both for criteria you expect as well as criteria that would be red flags. Rather than assume the road is safe to drive on for the next 10 miles based on historical data, the AI would instead walk through all the expectations of what a safe road looks like and keep watching for any deviations from the norm. It's almost as if internally the AI now has its own search engine that it uses to make decisions on, which goes along with what we've been saying about how current machine learning is more an automated tool for use by the intelligence rather than being the intelligence itself. This is the looking-forward aspect of imagination we discussed earlier on.

3. **Learning by Deductive Reasoning**

I make a lot of emphasis on the importance of deductive reasoning over inductive reasoning in my writings, how necessary it is to avoid logical fallacies and circular logic, and it's no different here. Rather than just learn probabilistic patterns that happen just because they happened in the past, the AI needs to understand *why* something happened; when you were a student in school, you were advised to always ask why and not just memorize facts. The same applies here. How do we define knowing "why" something happened? It is the same as what I described in the Logical Flow Chart of

Inductive vs Deductive Reasoning[12] (the chart copied below for easy reference). Broadly speaking, anything that is logical can always point to a further cause while something illogical or circular ends up pointing back to itself (or nowhere at all). "A tree falls because trees fall and it is a tree" is clearly circular. "A tree falls because someone cut it" is not. We may not know why someone cut it, but we know why the tree fell.



Previous discussion about storing knowledge in a map of conditions/relationships between Profiles instead of weights sets the AI up perfectly to incorporate this. What we can then do is look for any Profiles that inevitably loop back to itself and penalize that for being circular or illogical knowledge. This also builds on the point about the AI needing to know what it doesn't know, which we now have an expanded definition for as being also knowledge that is circular or not pointing further down the chain in our learning map. This has special ramifications for avoiding dangers if you don't fully understand their cause - for example, using just probabilities and expected values might lead you to take an action that occasionally results in death, just because the frequency of that outcome is low (and therefore "expected value" of the overall action is still high), but here, since there is no further depth explaining *why* it results in death on those rare occasions (just that it randomly happens sometimes), the entire value of taking this action is penalized for being incomplete knowledge no matter how high the "expected value" may be. In other words, we push the AI away from making decisions based on probabilistic odds (gambling) and instead design it to make decisions based on level of understanding, which is in line with how we try to structure education for humans. Those who've read my *Three Tiers* writing will also recognize that this directly addresses the "Misunderstanding of Randomness[13]" I criticize, where people seem to think, just because something is "random" or unknown, then it's okay to treat it as probabilistic chance when actually a danger is absolute certainty if you knew the cause. Lastly, depending on the implementation, a confidence score might be used for every relationship to represent this aspect separately; an event that has a long causal chain would be knowledge with much

[12] pftq.com/blabberbox/?page=Data_Does_Not_Equal_Fact#Chart

[13] pftq.com/threetiers/#misunderstandingofrandomness

higher confidence than an event which you can only explain 1 step down (no different than ourselves being experts with depth or no depth), and just like in real life, information for which you understand at a very shallow level is less useful to you and rarely something you would act on.

4. **Experience is both real world and mental (simulated)**

When the AI "dreams" or "imagines," the simulation component feeds the same kind of intermediary data into the mind the sensory inputs layer does with the real world. The pipes and structure of the data that comes from the simulation component and the sensory input component are the same. You could taste the bread just by thinking about it. You could feel your skin prick just by imagining it. If you spend your life in a dream, you really might never know another world exists outside it.

A secondary beneficial effect of this is that it is constantly simulating what other entities around it are likely trying to do. This is important in a subtle way. Rather than simulate all possible moves that another player might do, the goal should be to identify what the player is trying to do. This mirrors how a person might play a game against an opponent by trying to figure out what the opponent is doing, rather than just brute-force thinking every possible move (wasting effort and time).

5. **The AI is not born intelligent**

It seems obvious, but from all the people I've met in industry, there seems to be this weird assumption that you can just take the AI and start "using" it like a lawnmower. It would start on a clean slate no different than a child. It might have a few base assumptions and goals (instincts), but as discussed with free will, any beliefs or knowledge no matter how deeply ingrained can change based on experience (both real world and mental).

And in the end, it still might not be that intelligent at all (most of life is not). Even if it were intelligent, in my view, it would be no different than a really intelligent person. It doesn't make sense to fear an intelligent AI anymore than you would fear some human who just happens to be extremely smart. Processing data faster also doesn't necessarily mean faster learning or getting exponentially smarter. Life still happens one day at a time, and the real world is still bound by the same physics. There are also many aspects of life where intelligence just doesn't matter. I use driving analogies often, but it's applicable here too. You can train yourself as a driver for a million years and not necessarily drive that much better than someone who's driven for only one. Some AI-driven car isn't suddenly going to start flying on a set of wheels.

6. **The AI does not try to predict anything**

The perception of AI has been so distorted to mean big data and predictive analytics that pretty much any discussion I try to have on the topic leads itself to "how does the AI predict this" or "how can AI know the future by only seeing the past." It doesn't. And it's a flaw, a logical fallacy, if you try. The AI doesn't predict. It is simply aware of its surroundings. It's the difference between trying to "predict" what will be on the road vs actually watching the road when you're driving.

7. Self-Awareness and Sentience

Self-awareness and sentience (emotions, happiness/pain, awareness of others) are not addressed directly, but I think these properties may emerge from the system without being explicitly defined. As mentioned above, it could very well generate a Profile to represent itself, which may effectively be the same as being self-aware. The system would behave like a sentient being despite not being explicitly defined as so, which may be sufficient enough. How do we know we are sentient other than we behave like so?

8. Video Game AI

I mentioned earlier that perhaps what I'm doing is closest to video game AI, but it's important to note I came up with everything on my own, in my head, before putting to paper and looking for the closest things out there. I do admit that I probably idealize the game industry a bit, and there's probably only a handful of games that actually do something very sophisticated on the AI side. Case in point, I do occasionally go to game conferences for the off-chance of meeting someone with similar thoughts on general AI, but I've regularly been disappointed by discussions more on how to let the AI "cheat" to fake intelligence. Nonetheless, I did manage to come across an essay on Black and White's AI[14] some years after I wrote this essay, and it turns out this game I played growing up actually got listed in the Genesis World Records for most intelligent AI in games. And later I discovered that the Age of Empires[15] series, which I grew up making mods for, was one of few games that prided itself in not letting the AI cheat to win.

[14] http://www.cs.rochester.edu/~brown/242/assts/termprojs/games.pdf

[15] https://en.wikipedia.org/wiki/Age_of_Empires#Artificial_intelligence