Question 1

I thoroughly enjoyed Jeff's explanation of reference frames, and his ideas as to how cortical columns contribute to human cognition and behavior. I had previously encountered cortical columns as a major component in an explanation of Neocortical processes – in a talk called "Computational Meta-Psychology" by Joscha Bach, MIT Professor for Evolutionary Dynamics, at Chaos Communication Congress 2016.

In the talk, Joscha explained his "Conductor Theory of Consciousness". The role he assigned to cortical columns was a bit less complex: In his theory, cortical columns only represent "concepts", which could enter a true or false state based on the states of other linked cortical columns. For example, when I look at an apple, the column for the "apple" concept would enter the true state.

I saw this reflected in the Thousand Brains model: The voting mechanism which Hawkins uses to solve the binding problem would also assign such a state per column – the difference lies in the information that is provided by a column. For Hawkins, a cortical column's state does not just contain true or false: Instead, it provides what in machine learning theory I would call an embedding. I was quite surprised that Hawkins did not use the terms "embedding" or "embedding space" at all in his book. In Machine Learning theory, an embedding vector is a representation of a concept as a point within an arbitrarily complex vector space. Such a space is exactly what *A thousand brains* describes as "reference frames", in my understanding.

What further refines my previous understanding (if you can call it that) of cortical columns is the insight that they aren't just representing true or false. On top of that, they must provide a temporal model that correctly predicts sensorymotor input, since every cortical column is in fact connected to the sensorymotor system, according to Hawkins.

Beyond the insights though, Hawkins theory leaves many questions for me. For example, if a vote between cortical columns decides whether I see a vase or a face in the popular optical illusion, wouldn't that also mean that the competing columns for the respective concepts have learned independent, incompatible embedding spaces, so the embedding locations cannot be compared directly? Perhaps my understanding has some gaps, and I have jumped to conclusions somewhere.

Another point of similarity between Joscha's and Hawkin's ideas are their criteria for consciousness: A memory of attention, and a real-time world model. I am more and more coming around to accepting that consciousness is indeed somehow a phenomenon of real-time prediction in a PCA (Perception-Cognition-Action)-Loop when this loop is connected to a memory of attention. But this is hard to accept.

I find it even harder to accept that consciousness emerges from such a seemingly "simple" (actually not simple by any means) algorithm, when the roles of motivations and "old-brain" hormonal reward systems are dismissed as much Hawkins does. Joscha Bach described the coordination between cortical columns as the primary driver behind consciousness and linked it directly to Serotonin and Dopamine as the "currency" used by the brain to get cortical columns to learn stuff. But Hawkins completely dismisses them as having any role for intelligence whatsoever. At least I could not find a reference.

I would really like to know from him. What is motivation without Serotonin and Dopamine?

Question 2

In the following, I will try to construct an architecture of cognition for an autonomous agent based on the concept of cortical columns, strongly inspired by Jeff Hawkin's Theory of a Thousand Brains.

I really like the idea of active cortical columns as temporal models of a concept: While a certain column is in the true (active) state it must correctly predict the sensory inputs it is connected to. If it does not manage to do so, it may need to yield to another column which made a more correct prediction, and it will adjust itself towards making the correct prediction next time. I am not sure how I would implement that though, if not with some form of backpropagation? How does Hebbian Learning deal with an error signal?

There would also need to be a kind of selective exclusion between cortical columns. Perhaps the exclusion is based on the sensory inputs which a column claims to predict: Maybe some columns compete over the visual region where I see my cup of tea, while other columns compete for who gets to predict the region where I see my hands. Further exclusions need to be based on the motor neuron connections, so two neurons which perform different motions with my index finger don't emit conflicting instructions.

The goal of incorporating such cortical-column-ish ideas into a machine intelligence would of course be to create a universally capable and adaptive intelligent system. As Hawkins explained, such a system could not be created, without also imbuing it with a strong motivational system. But he left a large gap in explaining how the motivational system is actually connected to cortical columns.

The motivational system could be based on certain "goal columns", and the ability of other columns to predict the activation of a goal column as a consequence of their own activation. For example, you might try to reach the activation of a goal column which represents a concept such as "consumed sugar". This column would be given as part of the architecture. Other Columns would now need to be able to predict the activation of this goal column if they themselves are activated. While they are not activated though, their potential activation must be seen by other columns which facilitate steps along the way. For example, the "consumed sugar" column might be predicted by the "move hand with food to mouth" column. However, that one requires a "grasped food" column activation, and so on. This recursive evaluation of goals will eventually terminate and yield a sequence of actions to follow to reach the goal state. If the sequence of actions is followed and the goal column activates, the columns that facilitated the action sequence gain credibility as a "reward". This will make them more likely to be activated in the future.

I cannot however say how exactly the sensory prediction aspect of cortical columns would be tied into the motivational system. Of course, an active column must correctly predict sensory input. It must also make predictions about other columns which may be activated when they execute a motor sequence. Perhaps, a failure to predict either could lead not only to the activation of competing columns, but also to the activation of competing action sequences. In a sense, cortical columns would fight for participating in successful action sequences. A reached goal means a strengthened role for the participating columns.

Another architectural component that is strongly required is a memory of activated columns. This would be consulted when a goal is reached to distribute “rewards”, but also when an action plan fails (a predicted sequence of column activations does not occur). In this case, the recursive goal-action search would need to re-run while explicitly excluding or suppressing previously activated columns.

What concerns me however as a programmer is that the “useful” actions of a generally intelligent machine would always be a side effect of an abstract core goal. Because if the core goal is too direct, the machine would not develop general intelligence, as far as I can see. I suspect that goal design will not be as trivial as Hawkins likes to make us believe.