The roadmap to Artificial General Intelligence (AGI)

A. V. Grigorov

Abstract:
Intelligence involves language and linguistic reasoning; this makes it a phenomenon that is partly extrinsic to the brain, to the extent that language, as a system, is partly extrinsic to the brain. Consciousness and motivation/rewards (instincts/feelings) likely predate intelligence, from an evolutionary perspective, and ought to be treated separately. The following brief discussion on AI addresses those issues and proposes a plausible path to the development of AGI.

The actual roadmap to linguistically-reasoning, conscious AI is probably this:

1) Image recognition of verbs (not just nouns)

This should work in real-time from a video feed, not actual single still images. The easiest way to develop this might be:

a) Create a photorealistic simulated environment, which contains conventional 3D models of scenes and objects.
b) Teach a machine to do 3D surface fitting using a neural network (or multiple NNs), of those scenes and objects.
c) Up the game a notch and use 2D stills of scenes in the simulated environment to infer the 3D objects, processing and outputting the result using NN fitting of 3D surfaces like in b). This system can train on its own because everything is simulated - it can check how well it is doing by using the original conventional 3D model.
d) Make it all real-time, and progress to image/video recognition of the interactions between objects (this might need special hardware to be developed first).

2) Image recognition of cause-effect relationships between 'image'- recognizable words

-this is actually easy in the sense that correlation most often "is" causation. Correlation in this case is purely the chance that one "word" would happen shortly after another, e.g. "throw" precedes "flying object" most of the times. Any statistical software/code can do this. What this aims for is to create a map/look-up table of cause-and-effect relations between words.

On a small side-note, humans pick this up (learn to speak and think) by correlating events to the words pronounced by their parents. This is why a TV that is constantly 'on' would increase the risk of autism in a toddler; the words coming out of the TV do not correlate with the events in the room, and not learning the words (meaning) prevents learning the causal relations between the words and using those relations to think**.**

3) Implementation of concept causal map -based problem solving

- if a sentence defines a goal, tracing linguistic cause and effects, starting from the words in that sentence, should result in a set of actions to be undertaken to achieve the goal. Basically, what this means, is that the map/look-up table from 2) can be used to "reason" like humans do - linguistically.
For example:
goal: "have strong and light spear"
'Have' is causally connected to 'make' which is causally connected to 'assemble' which is connected to 'gather' which is connected to 'find'. 'Spear' is connected to 'long' and 'sharp' and 'at one end' (as properties, though we can also think about stretching the definition of "causality" a bit) , so you need to find long and sharp and light things. Then go forward again along the general line of causality (assemble it all).
Alternatively, 'have' is also related to 'steal', it's not like there is just one option.
'Have' is also related to 'receive' e.g. as a gift, but that doesn't lead to an action set. At least not to a very deterministic action set (you can't know if and when it would happen). Put more broadly, this would be the ability to run a symbol-based simulation/model of the world, vis-a-vis specific problems and situations.

This take on reasoning/intelligence naturally leads to an interesting definition of "truth" (as in "what is truth?"). "Truth" is when things relate with one another in the same way as their symbolic/linguistic/word counterparts.

4) Implementation of real-time 'immediate surroundings modelling'

- the ability to predict what would happen in 5 to 30 seconds (and more). If an object is falling - it's probably going to hit the ground. I am going to call this "basic consciousness". This is not just a mechanistic understanding of momentum and physics, but also, sort of, checking for the causal connections of the words that are "image recognized" in real time, to avoid surprises.

It is nonsensical to ask "what is subjective experience from an objective point of view". Subjective experiences are only definable subjectively. The objective point of view on subjective experiences deals with the physical implementation and evolutionary purposes, and not with the essence. This means that when it comes to consciousness, the essence is what you define it to be, introspectively.
I am defining consciousness very broadly as "subjective experience of the world". The question then becomes "what is subjective experience of the world?". This question naturally leads to the logical answer that consciousness is "having a running real-time simulated model of the world in your head".
It's useful to note that the model can be running without any sensory inputs, or with made-up sensory inputs ("what if" situations).

5) Implementation of ethics and running goals (if self-motivating, 'living entity', AI)

- This is the "why get out of bed" bit. For humans it's a whole hierarchy of goals and needs (survival, reproduction, well being etc.). For a "robot", this would be the main "goal" that is being solved in 3) including breaking it down into smaller sub-tasks and sub-actions as demonstrated in 3).

Ethics, in this context, would be hard-wired "no go" areas of the map in 2), in terms of actions/solutions and their foreseeable (by the AGI) consequences.
Evolutionary psychology speaking, hard-wired/unconscious/natural ethics allow for trust and cooperation in separate entities solving the same evolutionary problems (survival, reproduction, well-being etc.).

A simple motivation/running goal for a robot could be “thou shall obey humans”. This would make for a robot that spends all of its thinking time trying to find ways to obey humans.

Alternatively, a robot which is more human-like would have “thou shall aim for the procreation of entities like yourself”, as a running goal. The robot would then be likely to break this down into subtasks such as survival for long enough to procreate, figuring out what is “entities like yourself” in the absence of a definition for otherness, coming up with cooperative strategies if/when needed vis-à-vis other entities like itself aiming for the same goal, etc.

A broad ethics (“no-go” area) example would be “though shall not interfere negatively with the execution of the running goals of other entities like yourself”. A more specific example, in the context of the spear-obtaining example in 3) would be “thou shall not steal”, or “thou shall not have others being unhappy as a foreseeable consequence of your actions”.

It may be useful to make ethics hard-wired as opposed to emergent from the “main goal”. Emergent ethics can give rise to cooperative behaviour on their own, but they don’t necessarily guarantee trust between entities working on similar goals, because while the goals may be the same, individual circumstances and context would not be the same, which would result in different actions. Hard-wired ethics may improve efficiency, inasmuch they reduce the time needed to establish trust.

The above proposal for goals/ethics has conscious goals and conscious ethics. Humans, on the other hand, have subconscious goals (instincts), which are mediated to the conscious mind via “emotions” as a reward mechanism. It is possible to go down this route too, when designing AGI.

External stimuli and their processing is a continuum where at the very basic level the brain/body is just a transceiver (there is a 'feeling' signal), at the mid-level there may be some instinctive and hardwired response (both internal and external), and at the highest level the feeling lends itself to description and/or labelling. If it's describable, then it's thinkable, and if it's thinkable, then the world can be navigated with 'that' in mind (like with all other things). In purely mechanistic terms 'feelings' and 'emotions' are a hierarchy of variables, on top of which sits the 'happy-unhappy' spectrum. Seeking to maximize happiness is how goals are set and how the success of potential actions is assessed; all possible actions are assessed and rated depending how well you'd end up feeling after some indeterminate amount of time. If you felt nothing (and didn't have the concept) you'd be catatonic, in the sense that you may be conscious, but why do anything?
So, the implementation of "feelings" is actually quite mundane, ignoring for a moment the complexity and the hierarchy of emotions in humans, which reflect the complexity and hierarchy of the evolutionary goals of humans. Machine learning operates this way by default; there is always a variable that the algorithm seeks to maximize during the teaching/learning process, which determines the goal/success of the process.

Evolution and the actual developmental path that human intelligence has taken, have also made humans acquire emotional intelligence, which broadly speaking is the ability to juggle various emotions introspectively and combine them in a beneficial fashion. This can, in some cases, increase happiness without external actions.