Deep learning framework for Artificial General Intelligence - Part II

Shimon Komarovsky and Jack Haddad*

Abstract—

I. INTRODUCTION

DL as the..

from Order section of previous article..:

I think that the major issue in AI, which prevent it to reach AGI is the issue of will (or intention/meaning when it is sent to recipients). Humans always have some will, and it is like system's objective function tell it that there is a process with a destination. It is hidden in the need for big data and attention in DL models, as an implicit will that determine what should be the correct output. It is also hidden in in the curiosity-driven RL models, where curiosity is only one specific type of will, and in humans it is purposeful, i.e. specific. Similarly in exploration of RL. Instead, will should be addressed directly and explicitly.

One intuitive way to insert will, is simply externally in a specific input channel only design for that. This is the explicit way. Complementary or as an alternative, we can use communication and recent context.

Reminder about intuition behind attention: it is about how to calculate context. Instead of summarizing information flowing from input in a fixed manner, as in RNNs or CNNs, we perform selective summary of elements in the input (which is image patches in vision or a sequence of tokens in text), which determined via similarity to some query, hence a dynamic context. The queries are different in different NN architectures. E.g. in transformer's self-attention we query the elements themselves; in multi-modal the queries are intra-modal (self) and inter-modal (cross). This idea is implemented also in memory NNs, such as NTM [1] and DNC [2]. It is very similar to associations, since it based on similarity, and when several layers of attention exist, then it is like skipping from association to the other. This effect implemented also in multi-hop attention model. DNC include two important factors we use in our associative idea also: time of access/write and usage level, which may in our case be combined into some importance measure of the association elements. ME-LSTM model [3] also introduce the need of fusing LTM with STM.

I think that current meta-learning is wrong at the core, considering having shared parameters learned between tasks.

Because most of the skills we learn not related to previous tasks at all. Hence it should be progressive.

Logic is rigid. Therefore logic-based AGI such as cognitive architectures can act merely as inspiration, but not as means to construct AGI. AGI most probably should be based on neuro-sciences, since in order to handle a large variety of scenarios and properties that it should have - it must be flexible, fluid, adaptive, evolving and alike.
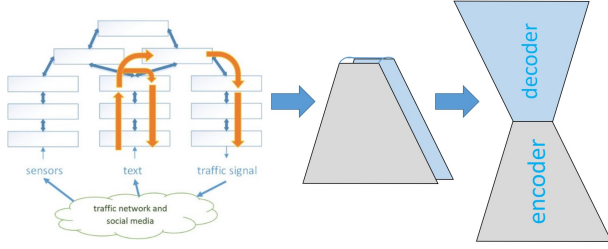
A. Models

1) General deep learning model: Inspired by the biology models and DLMs such as caption generation [4] and Visual-Question-Answer (VQA) [5], an Autoencoder DLM is considered. As shown in Fig. 1(a), the idea is to unwrap the percept-predict structure from [6] on the left, into discriminative-generative or encoder-decoder structure on the right.
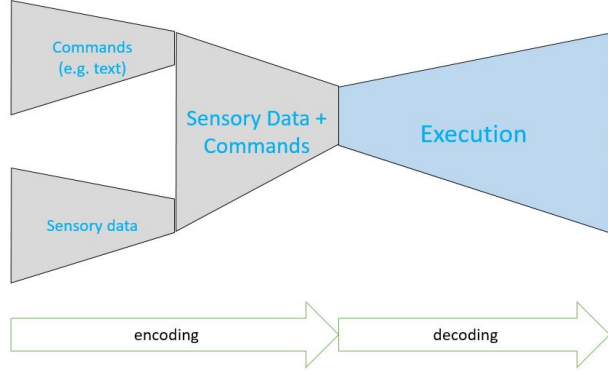
The more detailed proposed DLM is illustrated in Fig. 1(b). In this structure we encode all possible information, coming from text commands and sensors, finally into some extracted features that represent the whole situation including what the model is requested to do, and up-sample it to the actuators. Inside the model we have a situation description from sensors only (same process can be done for commands input), eventually combined into full situation grasping features, encoding the total information about what the model perceived and is asked to do.

There is an evidence about this multi-modal fusion in the literature of image captioning or video scene. E.g. a visual input is encoded into spatio-temporal space, and sentences describing the visual input are encoded into a continuous vector space. Then, the goal is to minimize the distance of the outputs of a deep visual model and a compositional language model in a joint space, and eventually to update these two models jointly [7]–[9].

The inner components of the proposed DLM are replaceable, and can be implemented via appropriate DNNs. Sensory input can be handled by CNN, DBN, SAE, and others. Commands input can be handled by RNN or its variants: LSTM or GRU (gated recurrent unit). The Sensors-and-Commands and the final decoder can also be implemented via sequence-based RNN, as in [10], since as we mentioned earlier, based on [6], the grasping of situation is gradual in time. It takes time to figure out the stable situation, and it takes time to follow up some desired plan. The plan is realized in a sequence of actions, such as in [11], where a robotic-complex-action is transformed, transferred, and used for the control of base actions.

Shimon Komarovsky and Jack Haddad are with the Technion Sustainable Mobility and Robust Transportation (T-SMART) Laboratory, Technion–Israel Institute of Technology. * Corresponding author: jh@technion.ac.il

(a) Percept-predict structure turns into discriminative-generative structure.

(b) Sensory data and commands in AE DLM structure.

Fig. 1. General sensory data and commands proposed DLM.

Note that the current model that was implemented, see Section **??**, is a short-range time dependent, however, for an efficient situation understanding we have to enable a long-range dependency as well, as in [12].

Our future research strives to develop the general model even further, to include perhaps a memory [13], either implicitly in the learning NN components themselves, or explicitly as additional components in the model, with different type of memories. E.g., a memory suited for the fusion of sensors with commands, which should hold different temporarily categorized concepts. I.e. fast changing concepts and slow changing concepts, and all in between. This conceptuality will enable the model extending its decoding for any kind of response, either textual or an action. In other words, we do not have a static type of memory, stacking up bits of data, but a functional memory. It remembers according to its user's use of it, i.e. according to the output.

This general proposed model is supported in other research topics, such as robotics and VQA, see more details in Appendix **??**.

2) Detailed deep learning model: In the future research, an enhanced DL model will be developed. It will be based primarily on two principles, in the following order: first we will extract features separately from sensors and text, then learn them together via joint embedding space. These fused features represent spatio-temporal information for the short-term temporal resolution. Next, we will implement hierarchical temporal structure of tasks [14]–[16], in order to implement scene understanding on different time scales (short, mid and long terms). I.e., the joint features will be

extracted further into longer time scales, by freezing first the short-term layers, and activating mid-term layers only. Same goes for the long-range layers afterwards.

Instead of only extracting features hierarchically temporally, we can insert in this temporal structure, an intermediate tasks that assist forming correct and more appropriate (guided) features, as in [17]–[20], where we recognize the relationships among the spatio-temporal objects we extracted previously. In our case, we would like to learn two types of information: objects and actions, starting from teaching basic elements/objects such as road, intersection, roundabout, pedestrian etc. Then continue to composite objects and actions.

The full sketch of the proposed model for future development is shown in Fig. 2. More details are given in Appendix **??**.
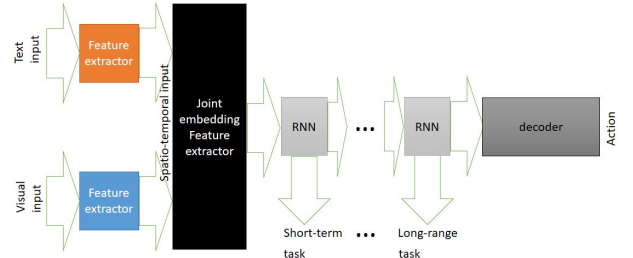

Fig. 2. A more detailed future plan DLM.

A more detailed future plan implementation is proposed, and mainly based on two ideas: How to combine sensory data with text, and how to apply hierarchical temporal structure of tasks, in order to implement scene understanding on different time scales (short, mid and long terms).

The first part is about extracting features separately from sensors and text, then learning them together via joint embedding space. These fused features represent spatio-temporal information for the short-term temporal resolution. In next phase of learning we extract these joint features further into longer time scales, by freezing first the short-term RNN layers, and activating mid-term layers only. Same goes for the long-range layers afterwards.

These hierarchical temporal RNNs can be implemented via different clock rates, as suggested in [14]. Other way is via sliding/shifted LSTM blocks as in [15], which are used to extract different time scaled features, and use those as an input to ST-CNN (instead of the usual short-term sequence of frames). Other way is via dilated casual convolution as in [16], which may capture also hourly, daily and weekly periodicity.

We can perhaps also test adding joint embedding space for spatial information only, before the spatio-temporal short-term joint embedding, as it is done with static visual images and simple objects in text [21].

Instead of only extracting features hierarchically temporally, we can insert in this temporal structure, an intermediate tasks that assist forming correct and more appropriate (guided) features, as in [17]–[20]. I.e. after the recognition of

objects in the features extracted in the two inputs, we should recognize their relationships. Hence the intermediate tasks are these relationships between objects. Some papers [8], [22] focus on pairwise interactions between perceived objects in an image, e.g. via 2D graph matrix, but [23] model high-order interactions between arbitrary subgroups of objects, using DNNs. We can try either of the approaches. In our case, we would like to learn two types of information: objects and actions, starting from teaching basic elements/objects such as road, intersection, roundabout, pedestrian etc. Then continue to composite objects and actions.

We should add attention units to different components in our DLM, as there are many types of attention implemented, such as described in [24]. For example, it can learn using attention mechanism as guidance: we input command="intersection" and in sensory input a graph/image with attention on some intersection, then we train it again on intersection with attention on other intersections. Similarly we can teach it instances of objects, such as "23th intersection" with attention on this specific intersection.

Similarly actions, e.g. "from-to", we show it a path from x to y, then we train it with different paths from same x to same y. And this can be done for different x's and y's. Also we can start from learning specific objects with specific NN components, then learning new stuff by adding more components on-top of them and freezing the first components if needed. E.g. learn about objects such as roads and intersections and other elements of transportation network including text and other resources, and then add new NN to learn new operations over these objects. Also, this system is reversible, i.e. we do not need to learn from component to new component only, we can train actions after objects they operate upon and recall that we forgot to teach some object. Hence we can retrain only the relevant component, and then return where we were previously. E.g., in [25] a gradual learning is proposed from a simple level to a complex level, either manually (expert guided) or automatically (scoring each sample by its training loss). However, this loss is highly dependent on the models and their hyper-parameters. Hence, a different learning takes place: from fewer categories or output tasks (local) to more categories (global). The full sketch of this model is shown in Fig. 2

Few additional thoughts present for the proposed model. First, we thought to have one channel for both text, which includes regular text input, describing the current situation, and the commands text - asking the system to perform something. But we should actually separate these channels, due to several reasons: (i) when we input commanding text, the NN trained by using executions as outputs. But what about descriptive text, what output would it yield in test time? (ii) It may be very often that both channels needed simultaneously, a descriptive one, such as coming from the user or from online social data sources such as WAZE, and commands one. Also we should remember that commands input represent our system's objective, i.e. an objective have to be defined always. And in NN the inputs are always active, so we cannot use an always active channel with two different

inputs. Descriptive text is also always on.

Second thought is about "dissipative intermediate tasks", meaning after "sticking" tasks to the outputs of hierarchical RNNs, we should train these RNNs without labels on the intermediate tasks we defined earlier, but instead on other tasks, perhaps in the action output tasks instead, in order for these tasks to be pined-out of the network, allowing perhaps the features in the RNNs become more generalized.

Fusion of sensors with commands: Note that in literature vision with text is used, instead of commands with sensory inputs. When we look at vision and text we have three types of tasks involving them. First type of tasks use visual feature extractors for categorical classification, which is a semantic type of output. There are also tasks that do more then one-word description of the visual features, such as caption description or language retrieval (choose best sentence describing a given video query). Second type is on the opposite direction, from text to vision. E.g. Word2VisualVec [7] which predict a visual feature representation (instead of a pixels) from textual input, to be used for language retrieval tasks, or video retrieval task (choose best video describing a given text query). Third type is the joint embedding of vision with text, such that it can be used for tasks for both ways. I also assume that these inputs are complementary. Since if they were trained together, then if one of them is missing, it is sufficient as if the second was there. E.g., [21] proved that joint embedding is efficient in Zero-Shot Learning, where unseen/untrained labels can be correctly predicted given appropriate hierarchical semantic graph. Vision features are considered as bottom, while Text features are considered as top, since the more abstract and more general features are textual comparing to visual.

For that reason, and the benefit of bi-directional inference, we concentrate on previous studies of the third type.

[26] demonstrates the combined input of image I and sentence S, for several output tasks, that scores the match of these input pairs. Tasks are confidence scores for answering an input question S, binary values indicating relevance of the input S for caption retrieval task, and binary variables specifying a set of image regions corresponding to the phrases in input S for the visual grounding task. [27] deals with the task of detecting several objects in a single image, and proposes for that a CNN-RNN framework to learn a joint image-label embedding, such that the embeddings of semantically similar labels are close to each other, and the embedding of each image should be close to that of its associated labels in the same space. The CNN extracts semantic representations from images and the RNN part models image/label relationship and label dependency.

[8], [9] propose joint embedding model of visual model that extracts semantic information from spatio-temporal images/video and of a compositional semantics language model which is a dependency-tree structure model that embeds sentence into a continuous vector space, by parsing sentence descriptions into ¡subject,verb,object¿ triplets. Based on this we can accomplish three tasks: video description, video retrieval (choose best video describing a given text query)

and language retrieval (choose best sentence describing a given video query).

In [9], assume word and image representations are first learned in their respective single modalities before being mapped into a jointly learned multimodal embedding space. It should be noted that in this research and others, hierarchy semantic structures are dynamic and constructed for a given phrase/sentence, which may imply that the real structure is not hierarchical (though it can be for feature extracting).

While most papers use minimization of distance between these two modals, [9] strives to maximize their inner products for the correct pairs and to minimize for the wrong ones. Unlike [8], [22] that focus on pairwise interactions between percepted objects in an image, [23] models high-order interactions between arbitrary subgroups of objects for video understanding. [23] also combines image coarse-grained (pooling from a sequence of visual extracted features) and fine-grained (high-order objects interactions) information for the action prediction task.

[28] deals with the scene graph generation task, where the output is a graph representing relationships between objects detected in a video input. It proposes the idea that it is not enough pairs of image and their descriptions for training scene understanding, but we also have to have some sort of memory or knowledge-base of the relationships between objects. Therefore, we support hierarchical learning as a way of constructing such knowledge-base.

3D incremental learning version of the model:

Until now we discussed the encoder part, but we can do gradual learning also including the decoder, meaning first we teach the encoder-decoder fast tasks with immediate execution, then fixing these first layers and then teach mid-term layers, and so on. This is the 2nd idea in incremental temporal learning. And it's a bit similar to how stacked AEs are constructed. This resembles the biology-based approach, where as for us, after repeating some task, it become automatic for us such that our mind is free from concentrating upon it, and now can deal with other tasks while performing these low-level tasks – same here, after accomplishing the low level tasks at high level of performance, we are free to "learn" new tasks. This potentially can introduce also working memory or thinking in higher available layers, to solve difficult tasks. We can view this idea better in 3d.

(a) 3D view of Fig. 2

(b) Fast stimuli-response layer

(c) Slow stimuli-response layer

(d) Slower stimuli-response layer
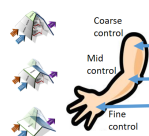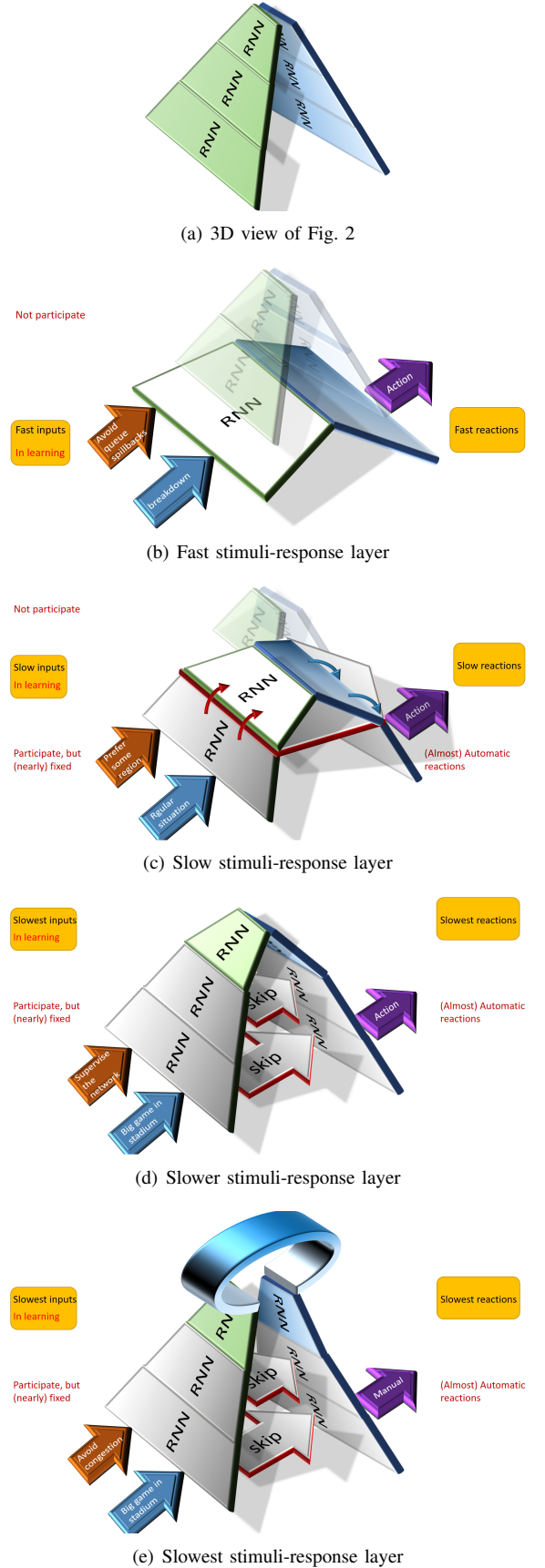

(e) Slowest stimuli-response layer

Fig. 3.   3D incremental learning version of the model in I-A.2.

In Fig. 3(b) we see fast commands. The input is the command and the state of the network. For instance: minimum delay in the signalized junctions, with the scenario of a breakdown, or another command of avoiding queue spillbacks, in a regular network state.

In Fig. 3(c), a slower commands for this moment are mainly preference commands. For example to prefer some region in a regular network situation, or not to prefer some region cause it's in a breakdown or under construction works. How it is related to the lower-layers which are already-trained and nearly-fixed? First, it allows recognizing some long-term patterns, such as of forming congestion or the effects of an accident somewhere, then, acting with some manner that changes the action space for the fast tasks in lower layers. Meaning, that when lower tasks were alone, they had the largest freedom or the least constraints on their actions, such as full range over green times. But when higher tasks are operating, they decide upon a more restrictive actions available for the lower task executions. We also have skip connections for the lower level to allow it continue operating almost independently. Specifically the connection that fast tasks had in previous slide.

Next layers, see Fig. 3(d,e), operate in the same fashion. We can imagine it might even be used for thinking, or solving difficult tasks, that are to be transformed to the more fine solvers in the lower layers. How is it possible? Because the inputs are very slow/stable, so these layers don't need to process the fast tasks, so they are left to deal with the other tasks given in the text input.

Implementation in traffic control problem:

Tasks	Description	Green duration	SP	Phase-Order
Slowest	Textual natural commands: "[input: there's game in stadium, [please handle it", "[input: coming important delegation in this route..], please prefer it"...	√	√	√ (assume objective is more important than effectiveness)
Slow	Preference commands: Green wave, prefer region (over), mail-function handling (e.g. incidents, cyber attack, power breakdown)	√	√ (assume SPs with different selection methods)	-
Fast	immediate response tasks: reduce spillbacks, min delay, max throughput	√ (assume changing demand)	-	-

Fig. 4. A more detailed future plan DLM.

There are plenty solutions for many cases, e.g. webster algorithm for min delay, for green wave for arterial, for specific events like signal plans for: And even some combinations of those plans. However, these are specific (discrete) set of solutions, i.e. for specific cases. We'd like to make it more continuous, to take into account also cases in between (Which is how DNNs or decision trees are good for). Or complicated cases, such as where we have conflict between solutions. Which allegedly can be solved by adding more complicated plans to the current ones - but it is exactly the deficiency of the rule-based approach, manually revise the discrete solutions, and add more and more to them. Examples are shown here. How SP works: for every intersection, we calculate fixed inter-green periods, then determine signal phasing and critical lane volume which will determine the green durations of each phase, which in turn will determine the cycle time. But this is for fixed SP. We can decide upon ranges for green durations assuming (slightly, since strong will require different SP) changing demand. We can also decide if the cycle time must be fixed (put constraint on the changing green times) or can be also in some range (put less constraints on green times). Note that critical lane volume is supplied by the data of demand (#vehs predicted to enter for every movement). Then signal phasing order is determined by the most efficient order for this critical volume, i.e. given the demand. So assuming critical lane volume is fixed at fast-task, i.e. it's not changing the phases order or their movements, only the allowable green durations and cycle time. I presume in slow-task we'll be able to override even the SP order, i.e. not even the effective SP selection methods (Approach phase, Protected left-turn phase, Overlap phase). I.e. can change duration time and which movements in a phase. The only thing it don't change is the inter-green times for security issues (and perhaps min green time for security also). Note that this partitioning to 3 levels, and different constraints on each are all prior knowledge. Be aware that in general this rule-based approach restrict the natural development of the model, self-evolving, to generalize beyond our specific prior, perhaps by encountering some data that we didn't include in our prior.. This is actually similar to "data snooping", is where the designer learns by himself, before the NN does, and apply this prior assumptions he learned to the NN, thus restricting the hypothesis space. Which might result better, but might also not. Anyway, the theory don't guarantee efficient learning in this case. Therefore we add as much as possible learnable priors, to be learned from the data and thus become posteriors.

Fig. 4

Examples: As it was in Tel aviv, there was power breakdown, some junction(s) were shut off, while other intersection continued work according their usual plans. A new project closed lanes in few roads in the city. Hence new plans must be designed. We'd like to have green wave but for some route of streets in a grid network. Conflicts (where we cannot choose one of plans above): Shabath and there's emergency (police or MADA). Work at Friday. Work and police (some emergency at working site). Covid19 require new plans (adaptive).
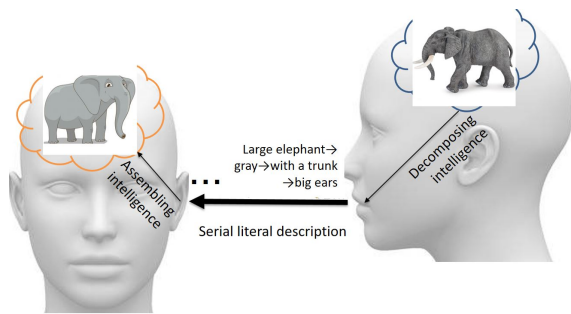
(a) Perception and communication
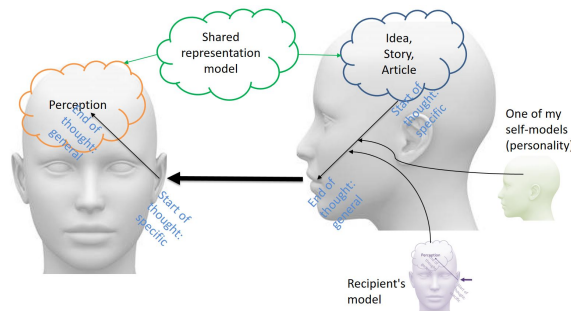
(b) Models in communication

Fig. 5. Communication basics

3) Communication: We recognize new prior to be added to DL models as we develop the DL subject, i.e. some elements that were taken for granted. Then human-AGI communication requires something more than only the recipients models (i.e. humans) - it requires that the AGI itself hold human-like cognitive properties and capabilities, so that humans and AGI would by synchronized during communication and understand each other. Hence it should have characteristics such as episodic memory, continual learning, abstraction, generalization, etc. Therefore, the AGI we trying to construct here can be defined as function-AI [29]. Though in general we could remove this prior, and let the AGI learn other agents' models simply by external tools, e.g. behavior.

This is related also with how much intelligence is a vague concept [29]. Since we may try to construct human-like AGI, and thus anticipate its behavior. But we cannot anticipate any other type of intelligence characteristics.

Important to note, that whatever AGI architecture we'll choose, we must be prepared to insert artificially, e.g. externally by learning, some of the prior embedded in human brain, perhaps inherited right from birth. For example, physical world priors, or the self and others models, and more. Moreover, some basic functionalities perhaps needed to be rule-based, i.e. written algorithmically, in the AGI, which are probably coming from the human genome, mainly regulating the main systems or memory. Similarly to **??**

Note that self model is needed for more than just communication. E.g. for controlling or changing own behavior and habits.

I guess many components or codelets in cognitive architectures, e.g. LIDA and MECA, together with brain development throughout all human's life span - is all embedded in the genome instructions in brain and other cells.

We are actually modeling everything, each thing - depending on our interaction with it. It applies for different people (different interactions), also different groups of people, similarly for each object or group of objects, and animals. Only interaction with human creates model by conversational interaction with them. We probably have also self-modeling, i.e. expectation from us, in the opposite direction of the interaction. How I should behave in different groups, different people, and different animals and objects.

Note that we have also a passive (active is above) "interaction" performed to model the entity in question, i.e. simple observation. Sometimes it can be as mimicking when observing other humans (e.g. parents or siblings). Also metacognition may be considered here as self-observation. Although it is not clear whether there is any observation on still objects. All the above describes the theory of embodiment.
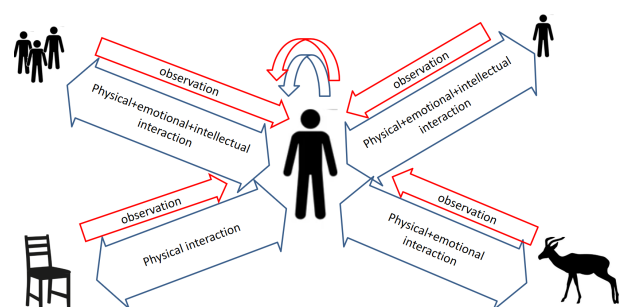
See Fig. 6.

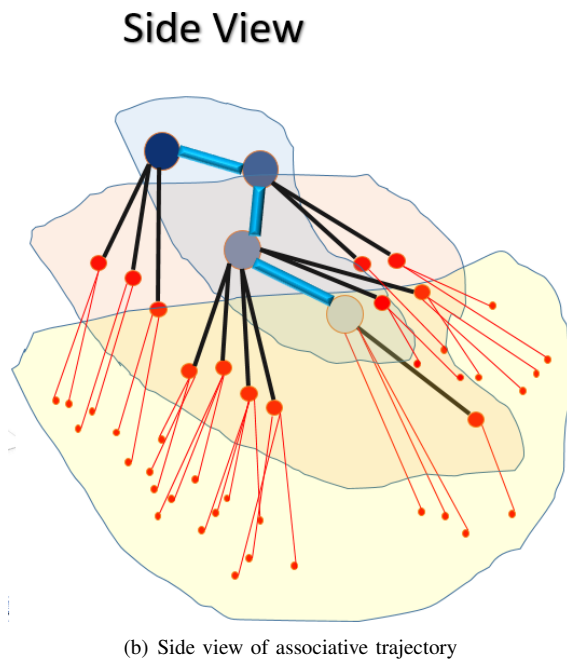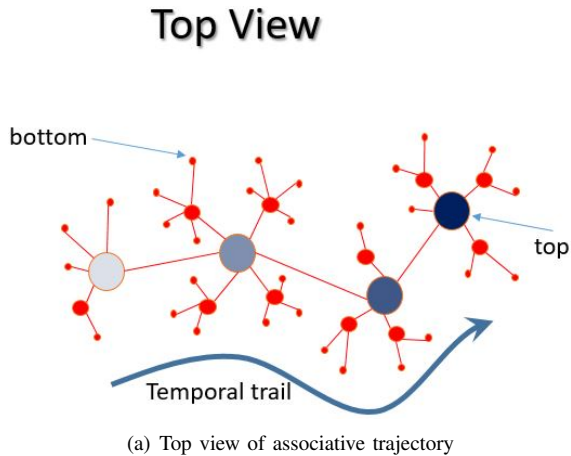Fig. 6. Human create models from interaction.

Top View

(a) Top view of associative trajectory

Side View

(b) Side view of associative trajectory

Fig. 7.   Associative thinking

4) Thinking: Associative thinking occurs all the time in my opinion. E.g., daily, the hierarchy is made up like a long story, with some experience at the top of the story's trajectory, made out of all separate events occurred during this day. But it can be attached to a previous hierarchy of previous day, and even previous week, month, year, etc.

It's important to note that this is a data representation model, not yet derived to the actual NN model to construct it. Additionally, this is a mature model, i.e. in adulthood, after there has been some stabilization. Hence this model also lack the evolution of memory till its mature state, i.e. all the primary learning and adaptation. It could be, for example, via self-supervising learning of predicting next sensory inputs. But this is of course limited to predictable sequences only. Most of them are doomed to be to complex for a verification of good data representation.

This idea kind of continue the principles of adaptivity, flexibility etc we have seen in dynamic NNs above.

Why it is important for contextual or associative hier-

archy? Because every element in a memory have many associations, so for any specific will or likewise - we need to distinguish it from other associations and hence different events and memories. So I figured out it to be as a high-level guiding function. In other words, we build hierarchies based on main concepts (at these hierarchies' bottom) to make huge amount of association distinct.

Important to note, that this is the output representation or its visualization, not the actual implementation. Next step is to perform some kind of an inverse engineering to figure out how this representation can be realized, in the form of some NN for example. For example, emerging thoughts in the WM can be implemented by some non-parametric method, since its structure is dynamic, similar to decision trees.

Associative thinking should in my opinion construct hierarchical trees, simply trees (where we deviate from the main story to some side notes/story for example), etc. Hence it simply about constructing trajectories or connections with additional data of the type of the connection (e.g. low/high-level connection, side note, and more). The hierarchy in my current opinion, is about different levels of will, from the most detailed aspects of it, to the most abstract or essence of the will at the top. Where top represent some kind of experience uniqueness or categorization, to differentiate it from other memories that use the same low-level structures.

Associative thinking explain communication structures as described above and episodic memory, and should include also other human-like properties. It should be similar method both for LTM and WM.

Other example that can be realized in this framework is problem solving. If we consider associative thinking as jumping from thing to thing, then in searching for a solution, or in the process of constructing one, we can view it as simple choice between BFS and DFS. The will is dynamic, and it will dictate when we go deeper (DFS) or when we go back to more preliminary nodes in our trajectory (BFS).

Moreover, will is dictating a subjective arguments. I.e. mind is an objective tool subjected to the will of the person. A person can justify about anything and its contrary - all depend on its explicit or implicit (sub-conscious) will/desire. Hence the bias and fairness which searchers battle with, is impossible to overcome, since there always be some bias, some preferred opinions over others, again - as a function of the will. Best thing we can do is externally change the will, which will yield different outcomes.

We can see the associative thinking as generating tree-structures on-the-fly, i.e. non-parametric approach. So it is a bit similar to dynamic NN we introduced earlier and others that exist in literature, but here they are constructed on-demand.

Inverse reinforcement learning (IRL) [30] or imitation learning is an RL variation that focuses on learning a reward function from human feedback. I argue that just as mirror neurons discovered by neurosciencesits, mimicking other peoples' behavior, similarly IRL and perceiving a story or a message from a sender (as described above) is all about constructing the actual intention from the observed behavior.

So infants, due to lack of developed memory must use visual (besides audio) input as a "story-telling" channel, to learn intentions of objects, either physical or living (e.g. animals and humans).

Evidence in neuroscience [31] is for example that concepts are not only from interactions, but also their purpose, which can be analogical to intention in our vision.

I claim that sparse attentive backtracking [32] is also simplification of the more general idea of associative thinking. Just as any human effort of isolating ideas from holistic systems, they end up be specific, and categorized as a narrow AI.

As in humans, system 0,1 and 2 is realized via this method also. I.e. when most frequent memory is used, it is for cases when automatic and no-thinking tasks performed, it is the system 0 and 1 equivalence. While thinking is for cases of solving problems and it activates LTM and WM. It also include cases where the system is fully utilized, i.e. we think but simultaneously also perform automatic tasks, which do not interfere with thinking.

Because there are many possible hierarchies in AGI (e.g. grouping of concepts, hierarchy in time, etc), then it is a problem to define which hierarchy actually exists in human brain, assuming there is one. One example of such a problem can be seen in the different hierarchies between the two books of Jeff Hawkins. In the first book [6] it is the regular DL hierarchy of features, while in the second book [33] it is about compositionality of objects.

As my great impression after reading the first book, I was totally convinced that it is temporal hierarchy, which is still a legitimate possibility. Since the current hierarchy I advocate for is actually very similar to temporal one - memory access. I.e. least frequently used memory is at the bottom, while the most used memory is at a higher level, while working memory as the current used memory is on top of that.

The smallest unit of AGI (atom like) is probably stimulus-response unit. Then when combined together as a network - it become huge stimulus-response block, similarly when combine many tiny magnets into a big one. Philosophically, I can argue that starting from infancy and up-to adulthood, there is a view claiming that we have the desire to control. But I propose a different view: stimulus-response basic element of AGI is the basic need for humans to fulfill their different wills in the environment. Hence, it yields the hypothesis of control. I.e. in order to accomplish what you want you need to learn how to manipulate your environment.

This notion is present in many papers of associative memory or associative NNs, where association is a response to a stimuli, which can be either other stimuli [34] or a behavioral response (action) [35] It can even fuse different modalities [36] All of above suggest, that associative connections should be extended into operation of some relevant concepts, where these concepts are attached to their relevant operations. Hence we can represent various associative connections. E.g. comparison, analogy, causality, correlation.

Also generalization, in the forms of analogy, metaphors, transfer learning and more - can be implemented as a special association connection between high-level trajectories. I.e. as additional level of connection. Generalization is interestingly can occur by somehow abstracting out different context and grouping the commonalities. For example, when i see dogs in different circumstances, and for each one of them being told that it is a dog, I connect all these events together, to learn some operational characterization if dogs: they have attributes like fur, small bodies, and especially their behavior is operationable. Similarly we learn math, by abstracting out the specifics of the examples we learn, left out eventually after lots of training, with an exact algorithm for doing math. Similar in any skill and action, we can generalize beyond some specific object, to perform the same series of actions over other objects as well. This may explain why we drawn naturally to rule-based approach.

I contemplate that we can model also an object or a concept via this abstracting also. For example, we can view a ball in different situations, and abstract out a higher level connecting them all, which represent the model of the ball. I.e. how the ball acts in different situations, e.g. it bounces of wall, also from a ceiling and the floor, it accelerates when thrown down or decelerates when thrown up, etc.

Notice that this dynamic graph-generating tool, can be used also for combining different models, or constructing new ones. I.e. allow for compositionality, a known approach in AI.

SHOULD THINK HOW ABSTRACT OCCURS WHEN KID PUT BALL IN BASKET, THEN THE BASKET MOVED TO OTHER ROOM AND BACK AND HE ASSUMES THE BALL IS INSIDE.. ALSO: WHAT WE BUILD IN OUR MINDS ARE MODELS OF STUFF, MORE ACCURETLY THE DIFFERENT ACTIONS/PROCESSES THEY CAN DO, LIKE BEHAVIOURS, EITHER SIMPLE OR COMPLEX.. I propose that the associative thinking is flexible and neural-based, but include also rule-based simple actions over the memories, like a supervising higher-level agent, that have very general tools at its disposal, e.g. analogy, relations, some very basic logic assumptions or transitions (e.g. if big don't fit, then bigger won't obviously), etc. But it must be very general and simple rules, otherwise they'll eventually end up constraining the intelligence evolving too much. I myself think that brain model shouldn't have symbolic operations, i.e. logic, because i consider these as external operations, things we communicate out to others or to the external world. But internally it probably is vague concepts, not well defined, to allow flexibility, and perhaps to allow associativity. Since associativity may occur on non-specific concepts or objects..

The question is how and where these operations stored? Perhaps embedded with their objects, or in a special memory. Also are there some basic inherited operations, beside the learned ones?

Anyway, we assume that memory data should not be in the form of merely knowledge graph or scene graph, since they are fixed and rigid, while we need more flexibility. In thinking we simulate different and new scenarios by applying

new actions on the data.

Its a system that constantly update the memory, i.e. storing new things and forgetting old ones. Similarly it store new tasks (stories, events, skills, thoughts) and forget old ones, because information is probably not separated into data and operations. In other words, data is actionable or operationable. It is a process. Also, the temporal sequencing in associative thinking allow to learn, plan and execute a series of actions. Note that we do not have forgetting mechanism, but it is a simple decay over time and in-usability. I.e. it is like a stack memory or LIFO, where new memories are piling up on top of the old ones.

Hence we actually remember processes (to do in future) or events and stories (past processes). Side note: this is why techniques turning random objects into memorable, do it by turning them into some consistent sequence or tale. We can say that the basic operation is as described in [37]: where system 1 is actually a multi agent system, where agents compete paralelly with each other to decide which pattern is percieved correctly from the senses, and hence which response is suitable to it. Similar idea in Jeff's, where this competition is via triggering all relevant neurons and filter out as more clue coming in from senses - all the irrelevant patterns, which predict worse than others, left eventually with the correct one pattern. Here we can see this idea, especially similar to on intelligence, by climbing up the triggered memorized hierarchies, and ascending down for prediction or verification, and hence filtering all the non-relevant memories. System 2 in my opinion is the cases when wrong top-down prediction require higher level adjustments in on intelligence, i.e. it is the learning level, which tackle new situations. After enough training and repeating of the same problem in system 2 level, it become automatic and decline down to system 1 fast memory.

This may explain why system 1 trigger relevant stories or narratives (parts of stories), to fill in missing information, or what's called common sense, which is implicit and not delivered information. Also partial or not-in-order info may trigger correct trajectories, by simple acsend in potential hierarchies, and after validate most appropriate one, being able to move in any direction of the trajectory. I.e. recalling some middle of a story, we can go back or forward in it, as we wish.

This may interpret Jeff's prediction of the next item in a sequence in a different more general light: it is simply recognizing the correct pattern, by examine it in different directions of the potential trajectories.

We can also view system 1 as effortless bottom-up triggering system, coming externally to us, while system 2 requires our effort, i.e. it involve our own will coming from above, so it is a top-down system in this sense.

Forgetting also necessary, since in humans it helps reasoning and understanding by forcing them to generalize and abstract [38]. It improves associations and relevancy, extracting the right patterns in a given situation, since it differentiate data to be scaled from being most common to least common. We can imagine it like data is hierarchical given this feature, which allow abstracting, verses flat data disregarding this feature, unable to make the data abstract or compositional.

Causality can be implemented in the form of ...

Compositionality, or how we combine functions, one after the other in a sequence, can be explained as ..

About creativity: as it is well known to be in right hemisphere, where the thinking is more divergent, the creativity come to play mostly when we are less looking for patterns and categorization. I.e. we can be biased or stuck in past patterns, when we find them, to find a solution to a problem. But sometimes (https://www.inc.com/jeff-haden/be-more-innovative-creative-better-solutions-research-how-to-solve-problems-more-effectively.html) we should dismiss the familiar in order to find novel ways to solve problems. In our case, it might be searching for associations in a BFS method and less in a DFS.

Note, that for constructing AGI framework we practice both the bottom-up approach (neuroscience, mainly based on Jeff Hawkins) and top-down approach (psychology, mainly based on Daniel Kahneman).

THINK HOW TO EXPLAIN YANN LECUN IN Q&A ABOUT STM UPDATING WORLD MODEL WHEN TELLING A STORY ABOUT GUY A IS IN A KITCHEN, HE'S EATING APPLE, GUB B IN GARDEN, THEN GUY B WENT TO THE KITCHEN. =¿ INFERENCE: GUY A AND B ARE NOW IN THE KITCHEN.

Associative model:

As a continuing developing models in I-A, we can summarize the model described thoroughly in I-A.4.

We can assume that simple sensory perception is using base memory, similar to system 0, automatic system (no thinking). Then it provokes LTM concepts or events, "uploading" it to the working memory. At sleeping period the system somehow decides what to consolidate into LTM and what not, due to unimportance or similar memories already exists. Note that LTM and working memory don't connect to the sensors and executions, perhaps since this is the abstract thinking, where thinking, depending on the task/will via some external instruction input, is moving in a purposeful trajectories, mostly or sometimes regardless to the inputs. This can be viewed in Fig. 8(a).

Actually there's permanent associative wandering in LTM, which produce some final or intermediate results which are updated in working memory. All this wandering must have some purpose as mentioned earlier, so there's some external will inserted in this process, guiding it. My believe is that we solve any situation/problem this way, by simply jumping associatively with some guiding will, which searching for something, gathering some intermediate insights, to eventually resolve with some response (good/no/bad solution). Actually we can regard the base memories, simple as part of the LTM, as the most frequent used in LTM, almost automatic in nature. Somehow we need to incorporate operational space of Jeff's 1000s... This can be viewed in Fig. 8(b).

(a) Discrete memories interaction.
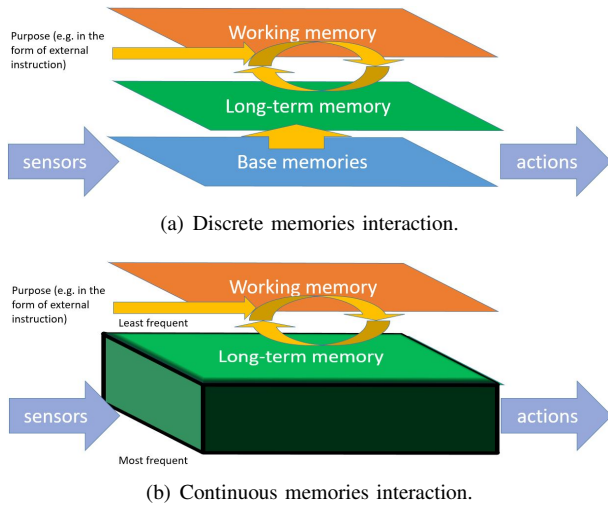

(b) Continuous memories interaction.

Fig. 8. Associative thinking model.

II. CONCLUSION

Most important takeaway from this paper, is the importance for holistic or complete attitude towards AGI. You should not construct it from parts.

REFERENCES

[1] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv preprint arXiv:1410.5401*, 2014.

[2] A. Graves, G. Wayne, M. Reynolds, T. Harley, I. Danihelka, A. Grabska-Barwińska, S. G. Colmenarejo, E. Grefenstette, T. Ramalho, J. Agapiou *et al.*, "Hybrid computing using a neural network with dynamic external memory," *Nature*, vol. 538, no. 7626, pp. 471–476, 2016.

[3] P. Liu, X. Qiu, and X. Huang, "Deep multi-task learning with shared memory," *arXiv preprint arXiv:1609.07222*, 2016.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[5] M. T. Desta, L. Chen, and T. Kornuta, "Object-based reasoning in vqa," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1814–1823.

[6] J. Hawkins and S. Blakeslee, *On intelligence: How a new understanding of the brain will lead to the creation of truly intelligent machines*. Macmillan, 2007.

[7] J. Dong, X. Li, and C. G. Snoek, "Predicting visual features from text for image and video caption retrieval," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3377–3388, 2018.

[8] R. Xu, C. Xiong, W. Chen, and J. J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework." in *AAAI*, vol. 5. Citeseer, 2015, p. 6.

[9] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014.

[10] W. U. Ahmad, K.-W. Chang, and H. Wang, "Context attentive document ranking and query suggestion," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 385–394.

[11] M. Suzuki and Y. Yoshida, "On the development and utility of action control individuality for semi-autonomous intelligent robots," in *2019 18th European Control Conference (ECC)*. IEEE, 2019, pp. 3550–3555.

[12] X. Wang, Y. Ma, Y. Wang, W. Jin, X. Wang, J. Tang, C. Jia, and J. Yu, "Traffic flow prediction via spatial temporal graph neural network," in *Proceedings of The Web Conference 2020*, 2020, pp. 1082–1092.

[13] J. Weston, S. Chopra, and A. Bordes, "Memory networks," *arXiv preprint arXiv:1410.3916*, 2014.

[14] K. Hwang and W. Sung, "Character-level language modeling with hierarchical recurrent neural networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5720–5724.

[15] X. Diao, X. Li, and C. Huang, "Multi-term attention networks for skeleton-based action recognition," *Applied Sciences*, vol. 10, no. 15, p. 5326, 2020.

[16] L. Ge, S. Li, Y. Wang, F. Chang, and K. Wu, "Global spatial-temporal graph convolutional network for urban traffic speed prediction," *Applied Sciences*, vol. 10, no. 4, p. 1509, 2020.

[17] D.-K. Nguyen and T. Okatani, "Multi-task learning of hierarchical vision-language representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 492–10 501.

[18] R. Cerri, R. C. Barros, and A. C. De Carvalho, "Hierarchical multi-label classification using local neural networks," *Journal of Computer and System Sciences*, vol. 80, no. 1, pp. 39–56, 2014.

[19] V. Sanh, T. Wolf, and S. Ruder, "A hierarchical multi-task approach for learning embeddings from semantic tasks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6949–6956.

[20] J. Wehrmann, R. Cerri, and R. Barros, "Hierarchical multi-label classification networks," in *International Conference on Machine Learning*, 2018, pp. 5075–5084.

[21] A. Li, Z. Lu, J. Guan, T. Xiang, L. Wang, and J.-R. Wen, "Transferrable feature and projection learning with class hierarchy for zero-shot learning," *International Journal of Computer Vision*, pp. 1–18, 2018.

[22] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1261–1270.

[23] C.-Y. Ma, A. Kadav, I. Melvin, Z. Kira, G. AlRegib, and H. Peter Graf, "Attend and interact: Higher-order object interactions for video understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6790–6800.

[24] Y. Yu, H. Ko, J. Choi, and G. Kim, "End-to-end concept word detection for video captioning, retrieval, and question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3165–3173.

[25] H. Cheng, D. Lian, B. Deng, S. Gao, T. Tan, and Y. Geng, "Local to global learning: Gradually adding classes for training deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4748–4756.

[26] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[27] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2285–2294.

[28] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1969–1978.

[29] P. Wang, "On defining artificial intelligence," *Journal of Artificial General Intelligence*, vol. 10, no. 2, pp. 1–37, 2019.

[30] S. Zhifei and E. M. Joo, "A survey of inverse reinforcement learning techniques," *International Journal of Intelligent Computing and Cybernetics*, 2012.

[31] R. A. Mason, R. A. Schumacher, and M. A. Just, "The neuroscience of advanced scientific concepts," *npj Science of Learning*, vol. 6, no. 1, pp. 1–12, 2021.

[32] N. R. Ke, A. Lamb, A. Goyal, C. Pal, and Y. Bengio, "Sparse attentive backtracking: Towards efficient credit assignment in recurrent networks."

[33] J. Hawkins, M. Lewis, M. Klukas, S. Purdy, and S. Ahmad, "A framework for intelligence and cortical function based on grid cells in the neocortex," *Frontiers in neural circuits*, vol. 12, p. 121, 2019.

[34] F. Shen, Q. Ouyang, W. Kasai, and O. Hasegawa, "A general associative memory based on self-organizing incremental neural network," *Neurocomputing*, vol. 104, pp. 57–71, 2013.

[35] M. U. Keysermann and P. A. Vargas, "Towards autonomous robots via an incremental clustering and associative learning architecture," *Cognitive Computation*, vol. 7, no. 4, pp. 414–433, 2015.

[36] K. Huang, X. Ma, R. Song, X. Rong, X. Tian, and Y. Li, "An autonomous developmental cognitive architecture based on incremental associative neural network with dynamic audiovisual fusion," *IEEE Access*, vol. 7, pp. 8789–8807, 2019.

[37] K. Daniel, "Thinking, fast and slow," 2017.

[38] B. A. Kuhl, N. M. Dudukovic, I. Kahn, and A. D. Wagner, "Decreased demands on cognitive control reveal the neural processing benefits of forgetting," *Nature neuroscience*, vol. 10, no. 7, pp. 908–914, 2007.