DEVELOPING AND MEASURING AI PLURALISM: STRATEGIES, MILESTONES, AND INNOVATIVE APPROACHES *

Devin Gonier
Columbia University
devin.gonier@university.edu

ABSTRACT

This paper presents a comprehensive overview of the development strategies and milestones necessary for realizing AI Pluralism...

***K*eywords** AI Pluralism · AGI · ASI · Multi-Agent Systems · Knowledge Graphs

1 Introduction

AI Pluralism represents a pivotal shift in the development of artificial intelligence, offering a pathway towards more individualized and adaptable AI agents. We define AI Pluralism as:

> *Multi-Agent frameworks in which A.I. agents interact with other human or AI agents, each of whom are diverse in perspectives and positions producing complex ecosystem of thought and recursive strategic refinement, as is manifested in evolution and culture for humanity.*

Many researchers have begun to explore how multi-agent systems might be configured using advanced LLMs. This transition from "mono-AI" to pluralistic AI represents an important paradigmatic shift in how one ought to approach doing AI research. Furthermore, it opens up the opportunities for other humanities based disciplines such as rhetoric, philosophy, game theory, anthropology, sociology, and psychology to contribute to the development of AI systems. The goal of this paper is to persuade the reader that not only is AI Pluralism valuable, but that it is also viable with the technology available today, and that with the right configuration it is possible to study such systems from many different perspectives.

AI Pluralism is a significant area of research because it offers great potential in alleviating many thorny issues in A.I. as outlined below.A

- **Safety** AI Pluralism can help to make AI systems safer by providing a diverse set of perspectives and positions and systems of self-accountability in which the ecosystem creates incentives for agents to keep other agents in line with ethical norms, much as trading partnerships enhance and stabilize peace between nations.
- **Bias and Fairness:** AI Pluralism can help to mitigate bias in AI systems by providing a diverse set of perspectives and positions. This can help to ensure that AI systems are fair and equitable.
- **Explainability:** AI Pluralism can help to make AI systems more explainable by tracking dialogical patterns in thinking as opposed to being restricted to opaque assessments of the systems internal state. The act of AI community building by its nature brings the internal outwards into a dialogical framework of external problem solving and contemplation.
- **Robustness:** AI Pluralism can help to make AI systems more robust by providing a diverse array of roles and abilities that enable coverage of weak spots through oversight. Much as the police make it their goal to identify and prosecute intruders, its possible to envision a system in which certain AI agents monitor the health and status of various systems.

* *Citation*: **Authors. Title. Pages.... DOI:000000/11111.**

- **Adaptability:** AI Pluralism can help to make AI systems more adaptable by providing a diverse set of perspectives and positions. This can help to ensure that AI systems are able to respond to changing circumstances and environments.

This paper argues has three main contentions. First, that the study of AI Pluralism is necessary for the pursuit of AGI and ASI. Second, AI Pluralism enables beneficial properties that are not possible with mono-AI systems. Third, AI Pluralism is a viable with technologies available today with the right setup, generative language mechanics, and retrieval systems.

The remainder of the paper consists of a deeper exploration of three components of AI Pluralism - the retrieval systems, the language generation architecture, and the situational context that are necessary for AI Pluralism to be viable. It then explores some specific experimental implementations of the proposed research and the current results of that research. Then, we discusses a broader framework for conceptualizing the relationship between AI pluralism and consciousness. Finally, we outline milestones for future research and conclude with remarks on how to accelerate development in this area.

2 The Retrieval Architecture

When humans think, there are at least two core processes involved. The first is the retrieval of information in memory, whether it be memories of events, concepts, or facts. What is retrieved must then be coordinated with the current context and objectives within that context, to produce thought as typically represented both internally (and often externally) in the form of language. This section outlines an expanded principle of retrieval in which features of identity or long-term states are retrieved and continuously incorporated into language generation. Current retrieval mechanisms such as RAG are useful for citing or justifying responses and mitigating hallucinations. However, databases can be used to do more than retrieve information, they can also be used to manage the state of the agent by defining and updating mechanisms such as beliefs, and interpersonal relationships.

2.1 LLMs Epistemic Primitives

Consciousness is not binary. It is not the case that one suddenly becomes conscious once the right set of attributes are in place. Instead, we should think of consciousness as scalar, which grows in proportion to the mechanisms that create useful complexity in the thinking agent. Some may argue for a more refined perspective, suggesting that consciousness is not binary, but it is quantic in nature. In other words, that there are mechanisms that when introduce result in large steps disproportionately such that qualities "emerge" as a result of interactions that have become newly available. Whether conceived of as quantic or scalar, we stand firmly against the idea that consciousness is a property that "appears" when the right conditions are met.

This suggests that there are components to consciousness, and that as components are introduced consciousness grows. As components are removed or reduced consciousness becomes reduced. For example, remembering historical events is a component (made up of other components as well) that contributes significantly to factors like one's perceived identity. Remembering historical events requires certain components to already be in place, and at the same time enables other components to exist if it is in place. This would suggest that a person who has a much better memory than other person may have relatively less consciousness. However, since consciousness is not a uni-variate system, it may also be the case (and often is) that the person with reduced memory skills makes up for it in other ways. I will refer to these components of consciousness going forward as **epistemic primitives**.

> Epistemic Primitives are the most basic components of consciousness that contribute to the complexity of thought that is possible for thinking agents. In other words, epistemic primitives are the building blocks of consciousness.

Epistemic Primitives should not be conceptualized in an arboreal fashion. In other words, while epistmic primitives may enhance or enable other epistemic primitives, we should not think not think about their relationships as a hierarchical set of dependencies. The trap with conceptualizing epistmic primitives in an arboreal fashion is the strong tendency to seek a fundamental root node, which does not exist. Instead, epistemic primitives should be thought of as a network of components that are interdependent and that can be added or removed in a variety of ways. A better metaphor for how epistimic primitves likely interact to form consciousness is how Deleuze describes rhizomatic structures in *A Thousand Plateaus*. There is no golden nugget, but instead a rich interplay of various components that make properties emerge as a result of scaling or level leaping (in a quantic framework)

3 The Generative Architecture

Focusing on the generative architecture, this part delves into Mixture of Experts (MoE) models, the importance of model merging, and strategies for routing and fine-tuning to achieve AI Pluralism's goals.

4 The Situation

Language games, as a measure of AI's progress towards Pluralism, are introduced here. This includes the examination of various games like 20 questions, murder mysteries, and debates, alongside a discussion on game theory and culture's impact on AI development.

5 Experimental Design

Our experimental design is based around studying how agents interact and cooperate in a language game context. Below is the basic setup for the experiment. We then provide our initial hypothesis for what to expect and what the results indicated.

5.1 Retrieval

Each agent in the system has tool based access to query an elasticsearch index to find materials that can help them answer whatever query or cirumstance they are encountering. These are broken down into Private, Semi-Private, and Public indexes.

Retrieval Accessibility

Private Index: This index is unique to each agent and contains information that is specific to that agent. This could be personal memories, or other information that is not shared with other agents. This private is also "inheritable" by future progeny of the agent.

Semi-Private Index: This index is shared with a subset of agents. This could be a group of agents that are working on a specific task, or agents that have a shared history. In the context of the situation we are studying agents form "tribes" and each tribe has a shared semi-private index.

Public Index: This index is shared with all agents in the system. This could be information that is common knowledge, or information that is shared with all agents.

Each index is further broken down into hierarchical categories and node types. The depth of each index is determined by the agent. Each agent is given access to functions which enable them to add documents to an index, or create a new index within an exising index. When querying, results can be a combination of documents and indexes. Thus, agents can choose to structure their index however they like in whatever organizational manner makes sense. There are some structural approaches we also intend on experimenting with where we remove some of the independent control of the agents. This includes basing index structure around node type, or document level. The document level follows the pattern of passage / passage-cluster / document / document-cluster / topic pattern. Since there are only three kinds of nodes in the index, its also worth exploring structuring the index according to node types: knowledge, beliefs, and relationships.

Node Attributes

Nodes in the index all have certain attributes, but there are some distinctions between them. In each index, there is a vector lookup which can be used with knn to construct clusters, and cosine similarity (or other similar measures) to find proximity between a query and index for neural semantic search, and finally the lookup vector can be used as a function in the graph neural network that sits on top of the entire network. GNN components are introduced in later versions however.

Another attribute every node type has is a set of edges. Edges define relationships with other nodes. Since every node in an index has an id, each node can track an adjacency matrix of its relationship to other nodes. These adjacency matrices can be further organized according to type of relationship.

Indexes are stored in s3 for tracking purposes at the end of each generation, and those which are not inherited are removed from the ec2 instance hosting the elasticsearch.

Node Types

There are three types of nodes in the retrieval network.

Knowledge: These nodes contain information that is factual or historical. They can be used to answer questions, or provide context for a situation. They are primarily populated with evidence found from internet search or other apis and always have a source attached to them.

Beliefs: These nodes contain information that is subjective or personal. They can be used to express opinions as they relate to claims and evidence. The key distinguishing feature here is a **credence** variable that doesnt simply retrieve information, but also provides a believability score. This is updated through the principles of Bayesian Epistemology. Beliefs may also represent a strategy of position on a particular topic.

Relationships: These nodes contain information that is relational or social. They can be used to connect agents, or provide context for a situation.

Each group (private, public, semi-private) can contain their own diversity of nodes (knowledge, belief, relationships).

5.2 Generative Modifications and Evolutionary modifications

Retrieval is essentially half of the influence on individuality in AI agents, the other half is the architecture of the models themselves. We achieve diversity in three different ways:

Fine-Tuning: Each agent has some aspect of it that is fine-tuned on a different source dataset. This could include how to perform a particular action or tool, or it might include domain knowledge. For example, in the language game of debate, one agent might specialize in a particular form of debate called kritik debating. This agent would be fine-tuned on how to construct kritik arguments.

Weight Merging: Each agent can also be further differentiated by representing a combination of a subset of the total pool of smaller expert agents. This is accomplished by selecting a base agent, and choosing 3 or more other agents to "mix" the weights together in a unique way. This kind of mixing has further variation in terms of the algorithms that are used for the merging, as well the distribution of the weighted sums for where and which layers to merge.

Retrieved Initialized Belief Structure: Each agent is fine-tuned at some subset of layers to incorporate beliefs into the system prompts such that those beliefs directly inform outputs. Arguments are always presented in the first person. The statements reflect the belief structure. As evidence and viewpoints are exchanged, these belief structures may evolve and change. However, the inital state can be diversely set.

These three methods enable a wide variety of potential agents to populate the situation. As each new generation of the game proceeds, further model mixing and initialization get inherited, identifying which techniques were most successful at performing well in the game and eventually converging on a community dynamic that is optimal for the language game task at hand.

5.3 Setup – The Evolutionary Language Game

A language game is any game where players must use language in non-deterministic and open-ended ways to make "moves" that lead to a win condition. Games must resolve in clear winners and losers. Every game has a "GameMaster" agent which ensures the game is played properly by the agents and keeps track objectively from the outside the state of each player in the game. Here are a few language game examples:

20 Questions: Agents are limited to 20 yes/no questions to guess a person, place or thing.

Murder Mystery: Agents are given identities and encouraged to communicate with one another and identify who they think the murderer is.

Debate: Two teams debate against one another, following academic debate models like Policy, Lincoln Douglas, Parli Debate from American Forensics. The Gamemaster's role is to function as the judge. The judge consists of a panel, which is itself diversified but optimized to not be in the minority decision. Note: that judge adaptation is crucial to be successful in this model!

Systems of Equation: Different groups are given different equations and must interact and communicate to figure out the hidden variables in a higher order equation.

5.3.1 Tribal Politics

In the first round of the game each agent can make a choice to either a) start a new tribe and invite other players into it or b) join a tribe that has already been started. Actions in the game cannot be taken by individuals, but are instead voted on by members of the tribe with capital. Agents can name their tribe, remove players from tribes, and invite new members into their tribe, but they cannot grow their tribe beyond a certain size. This ensures that there is a certan number of tribes to be consistent with the game mechanics of the game. This adds an additional game-theoretic element where the game is both competitive and cooperative. The goal is for the tribe to eventually evolve into a specialized unit such that different players on the team take on certain niche expertise that help the tribe succeed overall.

5.3.2 Capital Allocation

Each agent is initialized with a certain amount of capital which they can spend according to which actions they think are best. When proposing an action for a round, they can allocate capital based on their confidence level in that action. For example, in debate this might be a proposed argument. In 20 questions it might be a question and a justification for why that question ought to be asked. After all players have submited their questions, the tribe has a second round of spending capital where they select which action of the other players they want to spend capital on. The action with the most capital wins and the action is taken by the tribe. Agents submit reasons for why they think their action ought to be taken, essentially persuading the group of why their move is best. The agent who proposed the action, may get some capital back depending on the game mechanics of the game. If a player loses all their capital before the end of the game, they are eliminated from the game.

5.3.3 Evolution and Generational Inheritance

After a generation (a set of rounds of game play) the top scoring players get to "mate." Mating in this context means merging model weights with other winners. The amount a player gets to mate is directly proportional to their score. So the top player gets to mate the most, and the bottom player doesnt get to mate at all. The next generation then is made up of the progeny of the previous generation.

Furthermore, after mating, children inherit the retrieval indexes of their parent as a starting point, and the parent has an opportunity to update their beliefs one more time with a set of strategies for their children to utilize when they play the game in the next round.

6 Results

Here, the outcomes of the experimental setup are presented, providing insights into the practical applications and implications of AI Pluralism based on the conducted experiments.

7 Discussion

The discussion section reflects on the results, considering the role of language in consciousness as debated by Dennett and Chomsky, and evaluates the impact of culture and future research directions in AI Pluralism.

8 Milestones

This part of the paper proposes a set of milestones for measuring progress in AI Pluralism, including the complexity of situations, progress in language games, and sophistication of retrieval mechanisms.

9 Conclusion

The conclusion summarizes the paper's contributions to AI Pluralism, highlighting the potential for more diverse, adaptable, and innovative AI systems and suggesting avenues for future research.

Acknowledgments

This work was supported in part by...