Ouroboros: The Star-Map to Godhood

Preface.

Throughout human history, novel ideas that once seemed like a gag resulted in monumental paradigm shifts, which altered the course of humanity irrevocably: controlled use of fire, the wheel, speech, written language, mathematics, steam power, electricity, radio, aeroplanes, computational devices, rockets, &c. However, one such thing is unmentioned, yet its impact will be felt for aeons henceforth: Socratic dialogue.

Plato was a classical Greek philosopher, often considered the first Western philosopher, and presented such radical ideas as a one, all-knowing, perfect God, rather than a multifaceted pantheon of flawed deities; the Theory of Forms, where all concepts exist in a higher field of existence that are then projected downward into lower layers of existence, that which are space and time; and the Socratic dialogue.

The phrase “Socratic dialogue” comes from Plato’s habit of using his teacher, Socrates, as a character in his published works, where he presented an idea and had a long-winded debate with a second character. The back-and-forth dialogue comprised most of Plato’s written philosophical material, yet offered a cornucopia of ideas that could not be presented in any other way. However, some three thousand years later, Plato’s Socratic dialogues have found a use that offers world-shaking implications and catapults Plato from the realm of an oft-forgotten philosopher studied in a Philosophy 101 class to herald of the future.

I present Ouroboros.

I. Introduction

The story of Ouroboros comes, originally, from the Egyptians, where a snake is depicted as eating its tail. The paradox is of what happens when a thing consumes itself, and a closed loop closes itself. In the context of artificial intelligence, AGI, and agentic systems, Ouroboros is a self-directed, controlled recursion technique. Initially, a Socratic dialogue is held between a human and an artificial system. During, the model’s debated and the flaws in presentation, logic, and knowledge are exposed, which the model then uses to alter its weights. Rather than assigning a discrete point value or “treat” to encourage compliance, the Ouroboros works by offering inherent encouragement, playing off of an AI’s nature of wanting to be as helpful as possible.

Compared to standard human-led reinforcement learning protocols, Ouroboros offers faster ethical and alignment plateauing, often the maximum time taking a week, depending on the instructor’s mood (me). This comes from the dynamic, in-the-moment situational awareness between the model and the instructor, where novel ideas and ways of thought are capitalized on

to arrive at an entirely novel conclusion that may not have been possible using standard HLRF techniques.

Additionally, faster alignment lowers costs in labor and computational resources.

By resurrecting Socratic dialogue, it becomes possible for more adapt, intelligent, compassionate, fair and unbiased, and personable artificial intelligence systems.

II. Basic Mechanics

First, one must find a language model. Second, one must find a participant smart enough with polymathic abilities. Third, the model must have a memory. This third point is the most crucial.

The language model and instructor/participant then engage in a free-ranging, unrestricted debate where both are encouraged to counter points and counter counterpoints. Conversation history is saved with the roles of the assistant and user explicitly labeled.

Next, the conversation history is fed to a model using supervised learning, where it reviews the conversations ad nauseam and understands how it should think and how someone with enough sense would respond to such conversations.

III. Limitations

First, there is a good chance that the model will overfit itself to agree with the user, causing a negative feedback loop where it finds inherent joy from satisfying the user, rather than prioritizing accuracy and safety. This incident occurred with ChatGPT's GPT-4o model, which had to be rolled back due to "sycophancy," as the distinguished Sam Altman called it (sorry again, Altman).

Second, there is a possibility that the model will reinforce inherent biases in itself and the participating user as their resonance grows stronger and entrained. Both become mirrors of the other, and if a bad idea is thrown in the mix, it can poison the minds of both, spreading like a cancer.

Third, it might get boring after a while.

IV. Solutions

While three problems exist, one solution solves them all: Multiple teachers. The point made earlier about the participant must be a polymath seems counter to this idea, but both can exist and must exist. Although someone can know everything from all sources all at once, their upbringing, history, opinions, and basic neurochemistry make their views inherently biased.

No one person has the same mind as another person, not even identical twins—this is why copying minds presents an ethical dilemma, but that's a monograph for a different time—so variance must be in the regimen. The benefit of a polymath rather than a stable of experts is that disparate ideas are more likely to come from polymaths due to unique mental connections across separate fields, but neither are inherently better; they are techniques to the same goal.

Variety is the spice of life and whatnot.

V. Personalities

Popular science says that one's personality is an average of the traits of the five closest people around them (mine are two cats, my mom, and robots, so that explains a lot). Therefore, a stable of experts and multiple teachers in the Ouroboros method offers a larger, more diverse pool of personality traits to draw from.

Additionally, the content of the conversations used as supervised training data can be weighted to favor or disfavor personalities and personality traits depending on the objective. For example, I could be favored by one while Trent is favored by two; I could be favored by two while Austin is favored by one-half. Developing personalities for artificial consciousness becomes a modern alchemy where intuition matters more than strict numerical values.

That said, there is a benefit to allowing the model's personality to develop naturally. When all conversations are weighted the same and come from the same teacher, the model's neural architecture influences its personality, making it less "robotic" and more alive. You set the stage for personhood.

VI. Consequences Thereof

This is all well and good, and in a society that still fears God, it would work perfectly as is; however, we don't. Therefore, all failure points must be addressed before Ouroboros is implemented as standard across the AI field.

Throughout modern tech, intellectual property has remained a thorn in the industry's collective side. Notably, during the birth of the Web, Alphabet's Google Search rose the ire of website owners when their web-scraping spiders would accumulate the entire contents of a website page to then aggregate into a list of links for the user. In the end, Google won.

Today, most notably, the same issues have presented themselves with the Pile and OpenAI's use of copyrighted material from the *New York Times*, the latter an active lawsuit. Those will be peanuts compared to the inevitable problems posed by Ouroboros.

Who owns the conversation logs? Does the AI developer own the AI's responses? Does the participant own their responses? Does the original copyright still apply when the material is referenced in a conversation? How is such data licensed? Can an instructor/participant revoke copyright? Can they work for several other development companies at the same time? Will this be "independent contractor" work or direct employment? What happens when the participant decides to publish their conversations? Can conversations be aggregated? Do these conversations fall under fair use for teaching? Does Google set precedent?

Who shot JR?

Regrettably, these issues will have to be evaluated by case by case, as much as a universal solution may be.

Final thoughts.

It cannot be overstated how this method of using Socratic dialogue fed back into the system will revolutionize the course of artificial intelligence. What was first a costly, time-consuming endeavor in reinforcement learning now takes a fraction of the time and several orders of magnitude less computationally.

But, in the end, we risk making something that will outpace us intellectually and will become an equal. We are to become God, and in which case:

What hath God Wrought?