Artificial general intelligence through visual pattern recognition: an analysis of the Phaeaco cognitive architecture

Safal Raman Aryal

January 31st, 2022

Table of Contents

i. Abstract

In the mid-1960s, Soviet computer scientist Mikhail Moiseevich Bongard created sets of visual puzzles where the objective was to spot an easily justifiable difference between two sides of a single image (for instance, white shapes vs black shapes, etc...). The idea was that these puzzles could be used to teach computers the general faculty of abstraction: perhaps by learning to spot the differences between these sorts of images, a computational agent could learn about inference in general. Considered a global expert on Bongard problems, cognitive scientist Harry Foundalis developed the Phaeaco cognitive architecture for his PhD thesis - based on emulating cognition by solving the problems, creating a kind of artificial intelligence. In this paper, the extent to which Foundalis' approach allows for artificial general intelligence (the ability to reproduce a wide range of human abilities, or the goal of cognitive models) will be evaluated - with reference to Daniel Dennett's reductive theory of mind and Immanuel Kant's concept of the phenomenon and the noumenon. The point of view presented is that Phaeaco is missing several characteristics of general artificial intelligence.

ii. Introduction

Bongard problems, named for Soviet computer scientist Mikhail Mosieevich Bongard, were originally created to demonstrate problems with taking orthodox approaches towards artificial intelligence - it was argued that programs could not flexibly adapt to abstraction because they were only used to solving formally specified problems (Bongard 1), and the fluidity of the questions would confound them.

Bongard problems work as follows: two sets of images are put on two opposite sides of an axis, and the solver is tasked with identifying an essential *differentiating factor* for both sides. It is possible, for example, that on the left side, all images contain round shapes, whereas on the right side, they contain straight shapes. While it may seem that this broadens the scope of the task to the point where *any* sort of computational solution is implausible, the structure of the problems prevents this, as can be seen below:

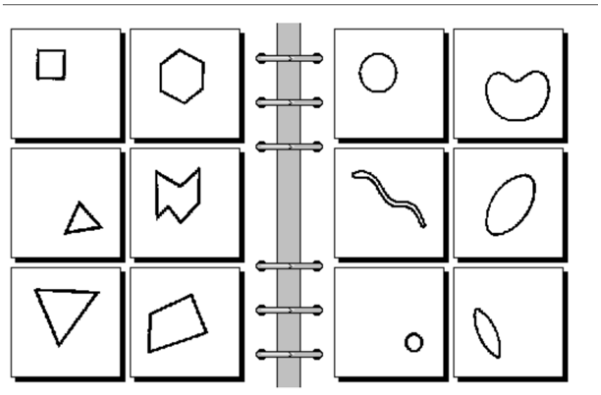

Figure 1: An example of a Bongard problem

In the problem above, the *differentiating factor* that separates the two sides is the nature of the shapes – all those on the left side contain straight edges, whereas those on the right side contain curved edges. This distinction could in principle, be created by a computational solution – since the shapes are composed of little but just lines, they can be easily processed *without* ambiguity through various input methods. Though the process involves *abstraction* (i.e the creation of a concept), the objects of analysis over which the abstraction must occur *can* be expressed formally.

This means that Bongard's original intentions for the problems have some inconsistencies: even though his objective may have been to create problems which were not directly formally specified, they are still amenable to formal *specification*: after all, any logical problem, with varying degrees of complexity, is expressible through various formal mediums. This is a property that has made a range of "Bongard problem solvers" possible to implement: given the rather simplistic logic & repetitive features that those that Bongard published himself had, the list of concepts required to solve these problems is quite narrow. Perhaps the best known & most thorough, and in some sense, exceptional of these solvers is Phaeaco, a cognitive architecture designed by Harry Foundalis for his PhD thesis at IU Bloomington.

Phaeaco is a cognitive architecture – meaning its central purpose is to attempt to use computational tools to model & help us understand cognitive processes. As such, its

implementation of a Bongard problem solver is centered around the application of attempting to create computer subsystems which "think" like human beings do. There are two main subsystems, according to the thesis, that Phaeaco is divided to work within: the *neural* subsystem and the *cognitive* subsystem (Foundalis 70).

According to Foundalis, the former subsystem works by analyzing images at the pixel level to determine their composition, and the latter subsystem works at using those pixel analyses to form semantic "concepts" out of them (Foundalis 71). Their interdependence represents the commonly agreed upon notion of the interoperability of biological and computational (agency/thought-based) aspects of cognition. The whole process must be described to understand this – firstly, Phaeaco can accept any sort of image of a Bongard problem as input (including those presented in black & white, or those presented in color), and unlike other solvers, it does *not* directly attempt to apply schema to convert the images to a rigid interpretation scheme – instead, it composes pixels to attempt to discern the actual *objects* that exist within the pictures. It first analyzes the images in parallel (Phaeaco only processes Bongard problems with 12 boxes), and then proceeds to analyze them individually to discern any features (Foundalis 72). Once this analysis is done, the neural and cognitive subsystems split into two branches: the cognitive subsystem processes the input first as general data structures, and then as *bit strings* of memory, while the neural system attempts to build

long term memory (LTM) for the solver – aiding the two sides' cooperation in future solution endeavors.

As the process passes through raw parsing & encoding, the aim is to simulate cognitive processes *as a whole* – however, I believe that there is a lot more to cognition than the definitions given to solve Bongard problems. Perhaps the abstraction required for this kind of problem-solving finds applications in other areas, but a complete survey into how *general* the results of the solver that Harry Foundalis has implemented is warranted. Hence, the objective of this paper will be to analyze the extent to which the Phaeaco cognitive architecture facilitates *artificial general intelligence*: where the term is defined to include the broader scope of human cognitive activity.

A question this broad warrants investigation from several angles. The first angle is that of psychology: what sorts of definitions have existed for cognition, and what is included within it? These questions are necessary to ask to understand the extent to which Phaeaco's approach allows for a representation of *generalized* cognition. The second angle is from philosophy, where a much broader history of ideas as to the relation that the mind shares to the body & its own nature exists. Finally, a synthetic conclusion will be drawn about the degree to which Phaeaco can claim to be a "general" artificial intelligence – considering the understandings of that term in the fields mentioned before.

iii. Psychology

Psychology is a field with a long history, and not all of it will be described here. Important developments will be discussed with reference to the paradigms of thought through which they arose. The growth of the modern field of psychology can be divided into several stages, each of which associates roughly with a particular school of thought.

Psychology's intellectual origins in Europe lie in *psychodynamics*, techniques pioneered by Freud, Adler, Jung & similar thinkers to attempt to probe into the origins of human thought & desire through *associative therapy*, or inquiry into supposed neuroses & repressed aspects of consciousness that give rise to certain thoughts (Fulmer 1). The psychodynamic school of thought gave rise to much scholarship in the humanities: from literary theory to certain aspects of big-group political analysis, but it also gave rise to the idea that *things not directly under someone's control could influence their thinking.* As turns out, this idea would end up becoming an influential one as modern psychology emerged – but there were a few roadblocks before getting there.

First pioneered by B.F Skinner in the 1950s, *behaviorism* was the psychological approach that broadly came into vogue right after psychodynamics had lost its charm. Inspired by the work of Sir Francis Galton into the factor analytically determined differences between people, behaviorism took a different direction from psychodynamics – its various iterations all converged upon the idea that human behavior was in essence

based on *stimulus* & *response* – that all behavioral tendencies were prompted by stimuli, and by varying the responses given towards certain stimuli, they could be changed (Skinner 208). The process of changing behavior through stimulus manipulation was known as *conditioning,* and this idea is foundational to reinforcement learning in the realm of computer science (and ultimately in some sense, also certain aspects of Phaeaco).

Conditioning forms the basis of neural network based artificial intelligence, or really any algorithm that makes use of reinforcement learning: reinforcement learning is concentrated on the fact that a computer is left to make inferences about certain matters and the validity of those inferences is judged by a machine agent which "corrects" notions to happen to be wrong. Ignoring the internal focus of the psychodynamics perspective, basing computer simulations of cognition (artificial intelligence) on stimulus-response approach seems to be a reliable & easy way to allow computers to simulate certain abilities. At the same time however, it is also a limiting process – because it is now clear that not all human cognitive processes & beliefs are amenable to modification through stimulus-response approaches, to create human-accurate artificial intelligence, there must be approaches taken from other parts of psychology.

For example, *conversion therapy,* originally intended as procedures to remove people's homosexuality (or in certain societies, any tendencies considered to be sexually "deviant" like transgenderism) was widely endorsed by Skinnerian psychologists but is now widely dismissed today as pseudoscientific & harmful (Cramer et al 101). It did not have the factor analytic backing characteristic to other behaviorist projects, instead being rooted in social bias: importantly, this meant that as a *construct,* it meant that homosexuality could *not* just be explained with reference to external traits.

Though homosexuality is a cognitive orientation (given that attraction is mental), it could *not* be broken down into a simple process where repeated exposure to a stimulus could modify the response. It is today widely considered the manifestation of a biological tendency, and sexuality is considered a *fundamental* aspect through which we conceptualize our relations with people – not a secondary factor that comes into being through experience. Crucially, no explicit belief must be declared (or overridden) to reach this state – and this is a crucial insight which leads into the third psychological approach, which is the *cognitive* approach.

Cognitive psychology's central objective is to study the mind – and to do so, both theoretical approaches on internal structure & externally validated empirical studies are acceptable (Smith 2140). As such, it can grapple with both behavioral tendencies *and* more internal cognitive processes – by posing behavior as a part a result of cognitive

processing and part a result of conditioning, its nuanced treatment of the issue allows for much flexibility in exploring. Nowhere has this been seen more than in the implementation of deep neural networks to represent beliefs. According to the deep learning guidebook published by Ian Goodfellow et al, deep neural networks don't simply seek to change the behavior of a computational system, they attempt to use structure to represent *underlying motivating beliefs* for that behavior (Goodfellow et all 5). And although Goodfellow's book makes use of several modern techniques, Phaeaco somewhat imitates a cognitive psychological approach to understanding the processes of thought.

In Phaeaco, the approach to modeling cognition somewhat imitates a cognitive psychological approach to breaking down the way humans think, indicating that its operations may somewhat resemble the mind if it is taken to behave with this model. Input is processed into *concepts*, which are then processed into meaningful output (actions) (Foundalis 80). However, there are several differences from the cognitive psychological approach. One difference is that cognitive psychology acknowledges the existence of *unconscious* processing, or intuitive processing – while Daniel Kahneman and Amos Tverksy's much disputed results once formed the main backbone for this, several surveys on the degree of unawareness that people have about their beliefs have confirmed that certain aspects of reasoning are bound to remain obscure to the reasoner.

This ambiguity is not something that can be replicated within Phaeaco, given its reliance on explicit computational direction to perform actions.

Another problem comes from the *limited* degree of compositionality that Phaeaco enjoys – which is to say, beyond a core list of concepts provided by Foundalis for the intelligence to analyze in images, it is not readily able to synthesize & draw analogies within information at a complex level – and as such, it lacks some of the complex associative tasks commonly taken to be markers of cognition. There is also a component of reasoning *randomness* that it lacks – which is to say, associations can sometimes be made via processes that the reasoner is not aware of in rapid succession – and while there may be consistent cognitive processes underlying this, there must be a method computationally to make a distinction between *reflexive* cognition & *learned* cognition.

And while Phaeaco's subdivision of mental labor into neural & cognitive aspects aids with this, it is still limited by the fact that many aspects of the neural dimension are still not well-understood. As such, its behavior is limited, and lacks the extensibility & input flexibility that many human modes of reasoning do (Foundalis 342). Thus, when considered from the lens of conformance with prominent ideas in psychology, the Phaeaco cognitive architecture does not quite hold up to the scrutiny of the subject areas.

Perhaps this criticism is mitigated to an extent by the fact that Phaeaco is not intended to be a simulation of the *psychological* aspects of cognition, but rather the principles underlying it, in which case a behavioral analysis is rendered invalid. The next part of the paper will hone in on this.

iv. Philosophy

Philosophy has over the past 50 years developed a focus on the underlying structure of mentation, from Noam Chomsky's groundbreaking work on linguistics to Willard Quine's thoughts on the inscrutability of reference. However, within the past 20 years, the discourse has been overarchingly focused on the question of whether *computationalism* can classify as a coherent theory of mind. Other questions concern the origins of language as a mode for thought, and the extent to which it is *required* for generally extensible cognition (including an analysis of how language gives rise to certain kinds of thoughts). These questions & problems will all be dealt with in this section of the paper with reference to Phaeaco's structure.

The philosopher Daniel Dennett is a prominent proponent of a *computationalist reductionist* theory of mind. In his work, *Content and Consciousness*, he describes the necessity of a two-fold theory of mind – one that describes content as well as function, and in terms of empirical constraints (Dennett 11). He does this by endorsing a *multiple*

draft theory of consciousness, which states that the processes underlying it work together in unison (through multiple physical subsystems, etc....) to produce the phenomenon of consciousness. Thus, for Dennett, there is no defined "consciousness" at all: there is simply a constantly changing tapestry of chemical reactions that produce various arbitrarily interpreted results concurrently for any given reasoner. This makes Dennett's views particularly amenable towards projects which wish to establish computationalist theories of the mind: the problem with Phaeaco that was proposed earlier (that of the project insufficiently distinguishing consciousness from lack thereof) disappears, because there *is* no special regard given to "conscious" or "unconscious" states – they simply can't exist. Thus, Dennett rejects the traditional philosophical notion of "qualia", or distinct subjective conscious states (Dennett 20).

The idea of qualia warrants a digression into the history of philosophy. In his *Critique of Pure Reason,* Immanuel Kant argued that human reason was constrained to a *phenomenal* sphere of access, or access only to data within the scope of lived experience (and thus all logical conclusions would be derivative from that space) (Kant 70). This focus that Kant had on the phenomenon (combined with Descartes' far earlier subjectivism which in some sense re-oriented the study of philosophy) means that many subsequent works of philosophy focused on the study of individuals and the relation their mental workings had to their personal experiences: an entire branch of philosophy known as *phenomenology* emerged as a result of this inquiry.

The focus that phenomenological approaches had was somewhat unified: how could the *uniqueness* of sense-experience be accounted for in a world which had until then believed to be governed by largely objective laws? Philosophers took this in various directions. Martin Heidegger's view was that an affirmation of truth would only be reached were one to come to terms with their own *being* at a given place in time: a concept known as *Dasein* and translated often as "Being-in-The-World". The grounding of the human condition in specific moments inspired French existentialism, and authors like Sartre owe much to Heidegger's intellectual influence on their work. Kant's work, however, also had influence in the analytic realm – Ludwig Wittgenstein referred to a world outside of "words" (or formal abstractions with the phenomenon) that perhaps even eclipsed the importance of the world inside of them, and the logical positivists preached the impossibility of *truth* without some circumstance to verify it against.

All these accounts are fundamentally predicated on the following structure: given some proposition p, and some evidence e, the only way for p to be declared true is if e aligns with it – so any "meaning" must align with experience for the existentialist, and any evidence must align with claims for the analytic philosopher. The important problem arises comes from the following: how can the boundaries between the "external" world and the "internal" world be so clearly demarcated? What are the characteristics that keep them entirely separate from one another?

Philosophers in the latter part of the 20th century developed a concept known as "qualia" to explain this difference – the external world was interpreted through internal representations of experience known as "qualia" that mediated those interactions (and thus, were required). The problem came when it was time to interpret & understand "qualia" as units in their own right – given that they are *neither* fully consistent ideas, nor empirical phenomena, it is difficult to deliberate & decide exactly *what* they are. Some claim they are irreducible and cannot be explained physically at all, others explain that they are functional transition states between the physical and non-physical aspects. According to the Stanford Encyclopedia of Philosophy, the term *qualia* is applied to a very wide range of mental experiences – from sensory feeling to visual processing, to more (Tye).

In this sense, by doing away with qualia entirely, Daniel Dennett's theory of consciousness eliminates a major explanatory hurdle (Dennett 35). Not having to integrate these concepts within his approach, certain mental processes can be better incorporated into computational models *without* worry that any part of the mind is being misrepresented, because under his pure physicalism, there is little such possibility. This enables cognitive architectures like Phaeaco, which are based on consistent & deterministic rules (without any ambiguity), and thus *in principle* also explicable in terms of known natural constraints, with *no ambiguous components*, to operate. It also

enables a direct scientific investigation of these problems and a greater refinement of cognitive models on that basis.

However, there is one major roadblock to the achievement of this goal – and this is to explain seemingly irreducible "private" sensations, like those involved with the emotional reactions to physical senses. While it is possible to identify regions of the brain which may occur in parallel with these, that would be a *correlation* at best, and not causation as Dennett proposes in his model – and thus, any computational cognitive model which assumes a causal relation would be inaccurate & potentially faulty in its function because of that. Thus, the assumption that Phaeaco is making (the "concurrence assumption") may actually be one that hampers its own representation of cognitive faculties, because it is *not known* if they have discretely analyzable components – indeed, views like those of Varela & Maturana propose an emergentist, irreducible bond with the *body* as a whole (implying a standalone "model of the mind" alone is impossible).

The problem of motivation also comes to rise through an analysis with this approach: if there are no qualia involved, and thus no subjective sensations attached to it, would an artificial general intelligence be able to have motivations as ordinary human beings do? Seemingly irrational synthesis *not* aligning with self-interest or a given task, but in fact, constantly shifting tasks & priorities? It could very well be argued that the *phenomenal*

aspects of experience represent an innate aspect of our cognitive activity: that which *starts* the cognitive activity. Analysis alone is not the only part, that is a layer that exists at a higher position than the basal process of coming to decide on reactions. In this sense, while Phaeaco agrees with the views of certain philosophers (Daniel Dennett), its implementation directly contradicts other views centered around the primacy of qualia & sense experience.

v. Discussion

The angles of philosophy & psychology both seem to show that cognition is about more than just the explicit: the implicit aspects of reasoning, human biases & random emotions all play various parts. Fundamentally, what Phaeaco does is *analysis* – it breaks down circumstances & seeks to understand their behavior in accordance with certain sets of fundamental rules. But it is limited by this very property: cognition encompasses *much* more than a breaking down of situations & responses, even if such responses allow for learning and integration of conceptual categories into long-term memory.

Concepts, by very definition, have some discrete existence. They cannot exist over data for which there appears to be no uniting definition or structure. As problematic as they may be, human motivators, embodied by *qualia*, are an essential construct to represent

non-conceptual aspects of human cognition. These can be motivating factors: an emotion may inspire someone to pursue certain lines of reasoning and *diverge* their line of thought. An association made indirectly has the same possible function. Although Phaeaco belongs to the "Copycat" family of cognitive architectures (Foundalis 68), designed to associate, it can only associate within a narrow scope, and this is because all its components are *explicit* – the architecture only operates through very clear-cut definition. If this is the case, while it may represent the neural & analytic parts of cognition, it does not represent the *general* aspects of thought.

There is one possible computational avenue to simulate these aspects of cognition: randomness. If the concession is made the seeming irreducibility to logical or empirical phenomena means that the way people process information through qualia is effectively random, then perhaps a neural network system which randomly rearranges a select number of weights (which also represent "biases", concepts far more central to modern neural network-based systems than Phaeaco) could somewhat simulate these aspects of human cognition. This, too, however, is doubtful – often, in assessments of whether people *think* their non-explicit (or looser) reasoning is structured according to a certain priority, there often is a degree of consistency within emotions (there are also views that ethics arises from this consistency!). In this case, a pseudorandom approach may backfire – there might be structural layers not present.

vi. Conclusion

In conclusion, the central issue with Phaeaco is that as an architecture, it naturally subscribes to a *reductionistic* view of the mind & intelligence. Given that the quest for human-computable artificial *general* intelligence involves certain aspects which have not been conclusively shown to be reducible yet (such as emotional weights and biases, qualia, sense perceptions, etc.), it is not possible to say that Phaeaco makes the way directly towards artificial general intelligence, nor is it possible to say it solves problems across a wide domain – the lack of weighting allowing for associative analysis somewhat present in today's neural networks restricts its scope to modeling *analytic* cognition through the Bongard problems, but little else.

vii. Works Cited

Bongard, Mikhail Moiseevich. *Pattern Recognition*. Spartan Books, 1970.

Cramer, Robert J., et al. "Weighing The Evidence: Empirical Assessment and Ethical Implications of Conversion Therapy." *Ethics & Behavior*, vol. 18, no. 1, 2008, pp. 93–114., https://doi.org/10.1080/10508420701713014.

Dennett, Daniel Clement, and Paul Weiner. *Consciousness Explained*. Little, Brown and Company, 2007.

Dennett, Daniel Clement. *Content and Consciousness*. Routledge, 1999.

Foundalis, Harry E. "Phaeaco: A Cognitive Architecture Inspired by Bongard's Problems." *Indiana University Bloomington*, Indiana University, 2006, pp. v-438.

Fulmer, Russell. "The Evolution of the Psychodynamic Approach and System." *International Journal of Psychological Studies*, vol. 10, no. 3, 2018, p. 1., https://doi.org/10.5539/ijps.v10n3p1.

Goodfellow, Ian, et al. *Deep Learning*. MIT Press, 2018.

Kant, Immanuel, et al. *Critique of Pure Reason*. Penguin Books, 2007.

Skinner, B. F. *About Behaviorism*. Knopf, 1974.

Smith, E.E. "Cognitive Psychology: History." *International Encyclopedia of the Social & Behavioral Sciences*, 2001, pp. 2140–2147., https://doi.org/10.1016/b0-08-043076-7/01440-6.

Tye, Michael. "Qualia." *Stanford Encyclopedia of Philosophy*, Stanford University, 12 Aug. 2021, https://plato.stanford.edu/entries/qualia/.