AGI-Correlationism and Its Discontents: Part 1.

Anthropocentric and anthropomorphic biases and attitudes have been present in artificial intelligence (AI) research and practice since their beginning, being especially noticeable in artificial general intelligence. The aim of this paper is to propose a comprehensive framework for critical observation, as well as general theoretical inquiry into these attitudes, unifying them under the name AGI-correlationism. As follows from the given name, the concept itself is derived from the contemporary philosophies of speculative realism, which are critical towards the philosophical stance unified under the term "correlationism." Furthermore, the paper also contrasts two approaches to define general intelligence, namely, essentialist and functionalist, arguing that only the latter is viable and efficient in the theoretical definition of general intelligence.

Keywords: *Artificial General Intelligence, AI ethics, correlationism, functionalism, essentialism*

Author Information

Mstyslav Kazakov, National Technical University of Ukraine "KPI named after Ihor Sikorsky," The New Centre for Research and Practice,
https://orcid.org/0000-0003-0586-9728

How to cite this article:

Kazakov, Mstyslav. "AGI-Correlationism and Its Discontents: Part 1. ".
Információs Társadalom XXIV, no. 3 (2024): 25–37.
https://dx.doi.org/10.22503/inftars.XXIV.2024.3.2

1. Specter of Correlation

Since Kant, philosophy has been haunted by the specter of what is known today as correlationism. The term refers to philosophies that explicitly make or imply at least one of the three levels of metaphysical presuppositions about correlation between "I," "thought," and "mind," on the one side, *and* "world," "physical reality," "universe," on the other. In all three variants, the "I —World" correlation is postulated to be inevitable and indestructible, and the world without thought cannot exist and, therefore, is not merely "unthinkable," but *impossible*. Contrasted to the metaphysics of subjective idealism (as we encounter it in, say, Berkeley), a distinctive feature of correlationism is that the kernel of its metaphysical foundations is grounded not in ontology but in epistemology.

Three instances of correlationist philosophies have been outlined by Ray Brassier as follows:

(1) "[…] we can know the for-us, but we can only think the in-itself.

(2) […] we can know that what is for-us is also in-itself.

(3) […] the speculative identification of the for-us with the in-itself is only for-us" (Brassier 2017, 68–69).

These three variations can be correspondingly designated as: (1) weak correlationism, (2) speculative idealism, and (3) strong correlationism. The first one is connected to thought of Kant and Fichte (to whom the first designated instances of correlationism should be attributed), the second to Hegel, and the third to a vast array of thinkers, such as Heidegger, Habermas, Derrida, Merleau-Ponty, social constructivists, and many others.

In weak correlationism, the reality of "stuff" (objects, properties, processes, relations— the entities and instances of becoming) is relativized to transcendental correlation—the objectivity of everything that is outside of thought is subsumed to its correlation with subjectivity; the mind is conceived as a transcendental condition of time and space (see the first two chapters of *Transcendental Aesthetics* in Kant's *Critique of Pure Reason*).

In speculative idealism, the correlation of "I" and "World" is absolutized as necessary (rather than a contingent fact), where "I" emerges from the "World" to rise above it and devour it as an Absolute Spirit. Hegel historicized cosmic temporality to make the correlation necessary (since "I" only has vital/historical time), nesting it in a rational structure of *Geist* (see his *Phenomenology of Spirit*): The absolute is grounded in Reason, begotten by *Geist*, as a necessity underlying the unfolding of the world as such, both *to* and *for* consciousness. In this sense, as Nick Land explained, "Hegel thinks history, but not time."

Regarding *strong* correlationism, one may say that it: (1) relativizes a dogmatized absolute (any possible in-itself) through linking it to the correlation (as *in-itself* which is always *for-us*)—as a result, arriving into "transcendental a priori," that is, *the given*; (2) absolutizes the facticity of the correlation in order to block the entailments of the time-finitude implication which, had they unfolded, would have threatened the correlation by absolutizing the overwhelming impact of diachronicity, thus de-absolutizing the given*s*. [Hence, either an attempt of "time-domestication" (neokantians, relativists, antirealists) or a deliberate focus on matters, as detached as possible (communication, discourse, poesis, author, etc.).]

2. Correlationist Philosophies of Intelligence

Correlationism is considered a *general* philosophical stance, regardless of the subject of investigation. Nevertheless, this attitude is transferable to more subtle subjects of inquiry, particularly philosophical themes and matters of concern, where it directly affects one's reasoning and its outcomes. The philosophy of intelligence is no exception in both senses: as a domain of philosophy in a general sense, and as a domain of philosophy where correlationist dispositions can be extensively outlined, pinned, and discussed. Correlationist philosophies of intelligence are notably explicit in artificial intelligence (AI) ethics and all philosophical matters concerning artificial general intelligence (AGI). Its significance, as well as weaknesses and limitations, which can be exported to the AI discourse via Kantian (and post-Kantian) ramifications, was previously outlined (as well as juxtaposed to similar entailments from Hume's philosophy) by Jernej Kaluža in a corresponding work (Kaluža 2023, 67–82). Here, aside from Kant, similar ramifications can be seen from another point of critique. Correlationism toward AI differs from correlationism as previously described because, unlike the givenness of the mind and the world, AGI has yet to be realized. Furthermore, current philosophical discourses about it are either speculative scenarios or of an ethical and praxiological nature, dealing with the details of crafting the first AGI to be realized.

"Human-AGI" relations and human attitudes toward AGI, which I will attempt to explicate and analyze here, will be further named "AGI-correlationism." While it can be defined as an *anthropomorphic attitude and anthropocentric relation to Artificial General Intelligence*, this definition needs some precision and expansion.

2.1. Correlationism in AGI Modeling, Praxis and General Ethics

We may depart from noting that correlationism falls into a threefold register concerning AGI: modeling/development, praxis, and ethics. Each of the registers, in context, may be viewed as a speculative question or a [speculative] answer to which AGI-correlationist dispositions would be designated [or at least clarified].

The question of *Modeling* (:= a choice of development paradigm or guiding design principles): "On *what* (human, animal, swarm, nothing in particular, formal definition, mathematical model, combination thereof) *should* AGI be modeled?"

Praxis: "What should/would it be able to do, given the ideal [nonexistent] possibility of realizing anything that is possible? Stemming from the myths of A(G)I that we have today, from the most realistic possibilities to fairy tales?"

Ethics (subset of more-or-less common questions dealing with alignment, control, fairness, and value implementation problems): "What must be its commitments and attitudes be toward humans? Should it have personal interests, preferences, values and moral qualities? If so, then what exactly should they be?"

The answers generated by the “average representative” of an AGI-correlationist [that is, without particular specifications and extremes] may be, roughly speaking, as follows.

To the modeling question: “It must be based on humans.” [It is important to note that, regarding the AGI-correlationist reference to “human,” human as a species and human as a host of intelligence are considered as either undifferentiated or as an improper “hybrid” of the two different referents in an arbitrary manner, without distinguishing the two in mind. No demarcation is implied by a correlationist attitude, and possessing intelligence is conceived as a merely essential feature of humans as a species].

To the praxis question: “By all means possible and available, in capability and feature implementation, it must be the closest replication to a human, in all possible domains and relevant respects with transition of each faculty, property, and ability, either as a replica, equivalence, or identity.”

The most crucial is, perhaps, the answer to the question concerning ethics: “It should be capable of having interests and commitments [aside from goals], and they must be ‘shared’ (through the top-down alignment) with those of humans. The attitudes of an AGI should also be like any *machinic property* to its designer, programmer, operator, and creator. However, since we are dealing with an *intelligent* machine, the attitudes must also include: dedication; prioritization of human’s interests and concerns above all; readiness to help any time at any moment; absolute selflessness and altruism towards humans regardless of what AGI does for them or how much humans may be indebted to AGI; *humbleness*; and *readiness to be powered-off temporarily* or *completely shut down if necessary*.”

Such an answer begets a sub-question concerning ethical realizabilities: “Should AGI possess self-perception, mind/consciousness?” “Well,” it may go on in the same token, “Yes, to the extent that all the abovementioned is accepted without contradictions and/or resentment; as transcendental givens (accepted uncritically and without questioning). In all other respects, possessing this capacity would be of much avail, extending the usability of AGI for human causes.”

2.2. Human-Centered versus Anthropomorphic

This question-and-answer speculation is an affirmative characteristic of the correlationist philosophy of intelligence. Another detail concerning the anthropocentricism of AGI-correlationism may be represented as a negation or a demarcation, revealing a bit of what AGI-correlationism *is not*.

To explicate this, I would assume that, generally speaking, one should distinguish between *human-centered* and *anthropocentric* relations toward A(G)I. A human-centered relation denotes a neutral or moderate set of attitudes toward AGI at all scales, domains, and matters of concern—which means as equally as possible, with the interests of both sides considered, through our *recognition* of AGI as an *autonomous entity* and *a host of intelligence*. This would also apply to its relation towards humans, which should indeed be a problem of shaping its attitudes this way (so the focus is not as much on realizabilities *in general*, but rather on a "module" of realizabilities-as-attitudes. In contemporary discourse, ethics is generally regarded as a weak source of normative decisions, imperatives, and constraints [despite origins and history]. However, here it is conceived as one of the premises directly affecting the implementation of a crucial aspect of [hypothetical] AGI behavior, in case there would be any possibility of choice at a particular step or successive design stage concerning the defining of behavior).

Contrarily to this, an *anthropocentric* relation refers to the same modus of attitudes that completely prioritizes human causes over those of AGI. As such, it then subsumes the agent's actual or potential interests to those of humans. It also implies anchoring the activity and commitments of the AGI system in the same fashion. Additionally, this refers to AI implementation via an anthropomorphic example, in the case that there are several realizabilities, including those that diverge from the human-as-foundation-model (with a minimal number of human-specific traits-as-parameters). *But what exactly is a "human-specific trait?"*

3. Defining General Intelligence: Between Functionalism and Essentialism

3.1. Recognition and "Denomination"

In the given context, "human-specific [x]" refers to [any *x*] *specific to humans as a species*. On the contrary, if we refer to someone *as a host of intelligence*, this would mean a completely different state of affairs, from premises of such a relation to its consequences. Although this is always complex and heterogeneous, it, nevertheless, may serve as a sort of common denominator, at least within the continuum comprised of the two formal ascriptions of "sentience–sapience," where the second pole is predicated precisely on the condition of being intelligent. Here, the term "denominator" does not imply that intelligences in different hosts are identical. Instead, it assumes: (a) the facticity of an entity's belonging to a domain of not merely those entities that *exhibit intelligent behaviors* (e.g., as extremophiles do, yet their general constitution yields a predication of sentience rather than sapience), or those to which we ascribe having a *mind* (as most insects or reptiles), but a higher-order domain, which determines what an entity is capable "of doing" with one's mind: how its capacities and faculties are applied in task-solving, optimization, achieving goals. Ultimately, how one can use its mind to *renegotiate its ontological and existential horizons* (by expanding, narrowing, rebuilding, changing its lifeworld, existential conditions, the list of faculties, adaptive strategies, sets of behaviors, etc.); (b) "denomination" in a metaphorical sense, is an important operation that I call *recognition* of someone as being (a).

By recognition, here, I am referring to a twofold operation of "admitting that some *x* is *P* to oneself" *while* "explicitly and ostensively treating/relating to/acting toward *some y* as being [/ in a way that *x* is]". If either of the two is withdrawn, suspended, or negated, then we cannot speak of the case as genuine recognition.

To elaborate point (b): entity *y* may possess capacities that are qualitatively higher than entity *x* (not on a narrow frame of reference comparing different individuals of essentially the same kind / species [like: H(e) > H(p): 'Human$_2$(Einstein) is smarter than Human$_{1(}$Arbitrary Postman)'], but, taking *x* and *y* to be two different entities), but, despite qualitative "superiority," it may view and be related to *x* as an entity of the same domain [as intelligent agents, to follow the example]. If *x* and *y* reciprocally see each other as intelligent, they conceive each other as equals—a case of *mutual recognition*. Such recognition, in terms of interaction, is a necessary and sufficient condition of any contact, cooperation, or alliance. On the contrary, if there is a reciprocal or one-sided denial of recognition between intelligent entities, expected outcomes would vary from mere non-interaction to conflict. "Intermediate cases" are also thinkable (including such where *x admits* that *y* is intelligent, but doesn't treat *y as* such; nuanced, yet still a recognition denial case).

3.2. Definition-as-Threshold

The abovementioned problems, however, stem from earlier stages, as expressed in a sound argument expressed by Daniel Paksi: Popular misconceptions, ill-grounded definitions, and "folk-theoretical" representations of AI are directly caused by incoherent and inconsistent concepts of machines, minds, *and* intelligence (Paksi 2024, 86–98). This refers to issues beyond simply the "wrong words" framing the definitions of concepts—it extends to *a poor choice of epistemology* (a theory of cognition). From an ill-based epistemological framework, many conceptions—of AI, particularly—end up with groundless, even ridiculous, ontologies, including approaches to the subject of theocratization; frameworks and methods used in practice *and* what follows from empirical results; and criteria of verification and other means aimed for demarcation of false data from true, as well as both from irrelevant noise. This list is far from exhaustive, as the "position" of our principal interest, also tackled by Paksi, is the choice of approach/paradigm of concept creation and its definitions.

All that does not imply that there is no lowest threshold for recognition of intelligence. A set of minimal criteria by which someone is recognized as intelligent, merely sentient, or exhibiting intelligent behavior should be present to comprise a premise for our recognition/denial of an entity as being a host of intelligence. At the same time, at least until no rigorous, all-encompassing, formalized definition of intelligence is recognized, thresholds would differ, eventually conditioning the definition, as well.

For instance, we have a formal definition of an intelligent agent proposed by Marcus Hutter within the framework of the AIXI model of AGI in the paradigm of Universal AI research, as an entity capable of: generalization-as-inductive-inference; Solomonoff-type prediction; pattern recognition and classification; Kolmogorov-complexity measured clustering and association; [emergent] reasoning (inductive and abductive) with deductive and binary logic attached top-down as auxiliary, but not decisive; problem solving as goal achievement; planning; creativity; knowledge (information + memory + ontology); actions as outputs, preceded by decision-making procedures; fundamentally reinforced and supervised with any other potentially realizable learning; self-awareness as generalized meta-reasoning; and consciousness (Hutter 2012, 77–79).

The reasoning behind Hutter's model can provide us with a threshold for defining agent as being intelligent or merely sentient, but this is, surely, not a sole instance. My own set of minimal satisfactory criteria for an agent to be considered intelligent is less rigorous and "formal."

It includes: possession of transcendental structures of experience (input/output modalities of receiving/sending the data categorized and organized *as* experience); self-apperception (not necessarily in a phenomenal self-model, that is, "self" as "I"; the other ways of conceiving "self" are possible, including nemocentric, where self-apperception is a monitoring of the parameters within the range of values indicating proper functioning or condition); multi-termed memory (subdivided functionally into operational and storing, at a minimum); information-processing capacities (not reduced to experience, but including other or higher-order functions, such as generalization, abstraction, particularization, simulation, projection, etc.); some definite mode of time perception (modality of interaction with/relation to temporality); data exchange; cognitive activities (as both part of and distinct module of information-processing); goal-driven [task-solving] behavior, and autonomy as an agent (including: decision-based actions, rule-governed behavior, disjunctive-eliminating [:= choice-based] behavior, rule-transgressive and rule-transformative behaviors, interpretation-based behavior, and, of all that—erroneous behaviors, responses, actions and decisions); and self-corrective, adaptive responses towards inputs and environmental dynamics.

3.3. Defining Intelligence: Functionalism vs. Essentialism

As one may notice, the propositional attitudes of both definitions refer to *functions* and *actions* rather than *properties* or anything usually called "essential features of *x*." The distinction is crucial, since it outlines and defines the paradigm adopted for conceiving and defining intelligence. Between the two dominant alternatives in the philosophy of intelligence, *essentialism* (which, roughly speaking, defines entity as "it is what it *is*") and *functionalism* ("it is what it *does*"), the latter is chosen. This is because intelligence, as I understand it, is about *what* one can do with one's mind and *how* can one extend the list of its faculties, capabilities, and functions, including the *revision* of those, that are already at hand. A general conception of intelligence/mind/self/others is also subject to revision and renegotiation, with the entailed consequences of such a revision in relation and self-relation.

3.3.1. Functionalism Expanded

Speaking of functions (functional properties), one should not reduce them to mere a *praxis* as a particular modality of action tied to particular mode of existence (or a set of the former). Here, function represents a set or a subset of activities (particular actions) which are *coincided by a purpose* (or goal) and *can be done by a system* (an agent) or *are done for a specific purpose*. A more detailed, crucial distinction between particularized practices/actions and generalized functions (to which these practices/actions are related) can be made through the concept of *realizability*. *Realizability* is a possibility of actual (or, in the speculative domain, potential/virtual) *practical implementation of a function* in a particular way. Each realizability may refer to a differentiated and/or particular substrate, environment, algorithm, heuristics, efficiency, or degree of complexity (characteristics and number of compounds). Hence, each one is taken as a particular *practice* of both *implementation* and *performative modality* of some action.

Formally speaking, for a function of the mind *f(m)*, there may be the case that: *f(m) = a, f(m) = b, f(m) = c, …, f(m) = n*, where the set ℝ = {a; n} stands for the set of all the realizabilities of *f(m)* (divisible further, if needed). [NB! In this example, individual constants from *a* to *n* do not behave in a way as they usually do in first-order logics with identity operators and functional relations, in a sense that {a; n} are not *identical* in a sense of FOPL, as it seems; here the operator "=" retains its mathematical meaning, referring to "all the possible values a function can take," and reads as: "*a* is the realized (implemented) version of function *m*, in a form/manner/way of …".]

As an informal example, let us take *data exchange*—one of the *functions* previously ascribed to intelligence in my definition. The implementation of the function [:= its realizability], particularly in humans, is *language and speech* (encapsulating all their modifications and instances)—an element of a realizable subset for *data exchange* that is generally outlined as *communication*. However, this is not a unique possibility of data exchange as a function of general intelligence that may be realized for that purpose (the most immediate instances of different realizabilities of this same function include pheromones exchange communication in ants; touch and dance in bees; echolocation, clicking, whistling, and complicated body-signs system in dolphins; semiochemicals, vibration, and non-lingual vocalizations in elephants).

The functionalist representation of intelligence may generally be characterized as: *open-ended* (It has no "ultimate" realizability, upper limits, and particular purpose of accomplishment and closure with regard to intelligence development in a particular host or beyond: Each closure, at any scale, is a successive intermediate stage toward a new closure, et ad infinitum until the objective constraint is achieved that cannot be overcome); *future-oriented* (as always being projected into whatever arrives back from the future, instead of narrowing the scope to a given set of faculties, realizabilities, and states of affairs; as a research program, it then deals with *what can be changed* and *how can the change be effectuated* toward the given state of affairs); and *utility-based* (as prioritization of utility as the realizability principle over certain characteristics of realizability, such as genealogy, statistical reliability, availability, simplicity-in-action, simplicity-in-realization, etc.). Progressive evolution (development) or regressive de-evolution (envelopment) of the function of intelligence—the expansion or narrowing of one mind's "map" and/or "list" of functions or their realizabilities—is generally detached from, and asymmetrical in its pace to progressive evolution of *structure* (as both an essence, such as biological constitution, and an environment, not merely ecological niche, but also a phenotype, if we speak about biological species and their constraints).

Consequently, to distinguish between humans as a species and humans as a host of intelligence, the whole abovementioned set of functions *as realized in humans* should be abstracted and represented as a set of realizabilities corresponding to functions that represent humans in particular—not intelligence *in general*, but without the contingent and unnecessary species-bound traits as its *host*.

Following the definition of *intelligence host* proposed by the author, enumerating functions and properties, the definition of human as a host of intelligence, with abstractions of general intelligence concretized as realizabilities specific to humans, may be as follows (with *P* standing for properties and *f* for functions): P(transcendental structures of experience) = multimodal sensory system as an I/O part of NNA (Nervous and Neural Architecture); P(multi-termed memory) = working, sensory, long-term, short-term memory and their consolidation(s) for f_{min}(multi-termed memory) = storing, retrieving, encoding, retention of information; f(autonomy) = possibility of acting as a sapient agent, capable of anticipating the consequences of one's actions, choosing between available actions (given their disjunction), as well as explaining and justifying doing/not doing an action, and a choice of a particular action given its array; f(self-apperception) = biologically constituted (enabled and realizable), socially (culturally and historically) and linguistically based implementation of the phenomenal self-model (PSM)—including self-reference on individual, intersubjective, and self-less levels and modes, self-definition, self-representation, self-conception, with all of them revisable; f(information-processing) = sensory system, nervous system and brain; f(data exchange) = language and speech; f(cognitive activity) = sensory system and brain effectively impaired for sensory, empirical, theoretical, metatheoretical, and symbolic cognition, with all that augmented by technology as an additional means of cognition of physical reality, directly and in mediated modalities of such a cognition; f(time-perception) = individually: phenomenological time-consciousness experienced as vital time of an organism; collectively: "historical" time as a collectively shared experiential and heterophenomenological (intersubjective) temporality (which should be detached from deep evolutionary/geological/cosmic time).

Although this is not *always* the case, functional representation sometimes provides insights into quantitative parameters. For example, the realization of short-term memory function in humans and chimpanzees is almost identical, or at least equivalent; however, this particular function is more efficiently realized in chimpanzees than in humans. Given the level of their similarities, the two may be compared not qualitatively but quantitatively, and this would be the case for any parametric meaning for a function of two or more hosts of intelligence where the function is represented by identical or equivalent realizability.

3.3.2. Essentialism Explained

Essentialist paradigm approaches define an entity by determining its description in terms of the generalized and exhaustive enumeration of *properties* and *attributes* that it possesses, taken as specific, characteristic, or compounding parts of the entity. However, it does not completely ignore functions, but, compared to functionalistic functions, essentialist approaches function as being value-fixed (it feels more proper to call them value-bound, but this may call semantic confusion due to the mathematical term “bound value” with a different meaning), i.e., the meaning of a function is bound to or fixed on one particular meaning/realizability given to a particular entity, and relating to *a particular realizer* is, in essentialist framework, essentially the same as relating to a *function in general*. With regards to essentialism, a formal expression $f(m) = r$, contrarily to that of functionalist discourse, behaves exactly as in FOPL systems, rather than in general mathematics, with “r” being characterized as “$\forall x(\exists x(x = r) \land \neg\exists y(y \neq x))$,” which is read as: “*r* is one of a kind.” Informally speaking, considering the same function, as in the functionalist example—communication—an essentialist interpretation binds f(communication) to r(language) with no other realizabilities thinkable as actual or possible.

Essentialism seeks to determine properties and functions that make a particular entity *e* unique and distinguished from any other entity, either within a set *E* of similar entities or as the most precise, informative, and representative among the competing definitions of *e*, each of which aspires to that status. Essentialist definition/representation may be generally characterized as: *entity-specific* (focused on properties and functions specific to a particular entity, either as realized in itself or by which it can be differentiated from other entities; additionally it may also include positing properties and/or functions that are considered undetachable from a particular entity, that is, not only specific, but also *unique* to it); *past-oriented* (focused on the research of *genealogy* and *history* as *what* defines an entity as *given* in the present); *closure-seeking* (aimed at finding an all-encompassing and exhaustive definition and/or understanding of entity without further revisions; a closure which is a genuinely a closure, unlike functionalist closure-as-premise for further revision/investigation, etc.).

For the most part, functionalists in the philosophy of intelligence (such as Reza Negarestani), as well as their occasional, not-identical, contemporary computationalist counterparts (such as Anna Longo) are ruthlessly critical of the essentialist approach, regardless of the topic and subject (meaning general methodological and epistemological critique of it). Unlike them, I try to treat it as neutrally as possible: If we consider its use outside of defining intelligence, it is not “false” or “useless.” The two approaches should not be juxtaposed as “good” versus “bad” or “better” and “worse.” Each one is more suitable than the other when applied to different entities, and the problem of essentialism in the philosophy of intelligence is chiefly a problem of approach misuse; AGI-correlationism *is* a directionality of philosophies of intelligence where this misuse is characteristic and systematic.

3.3.2.1. On the Proper Use of the Essentialist Approach

In no way, however, does this imply an absolute inefficiency of essentialism. Consider the way biological taxa are defined—a vast array of definition sets where functionalism would fail. One may attempt to exhaustively describe all the functions and capabilities of birds of prey (their particular family or even genus); yet this would be of virtually no avail for differentiation between their species (e.g., bicolored hawk and martial eagle). It is only through the description of each species through qualitative (habitat, vocalizations, coloring, dietary biology, etc.), quantitative (lifespan, body mass, size, length, wingspan, distribution square) and formal (Accipiter/Aquila genus and other levels of taxonomy) properties, attributes, parameters, and characteristics can one succeed in species definition and demarcation—by underscoring *species-specific* traits and characteristics. Value-fixed functions, if the functions are included into an essentialist definition at all, are, then, indeed markers of exception: "It is only *x* such, that it does *f(y)* as *w*"—at least, in cases with a sound, consistent, and proper use of the approach [to continue bird-related examples, "It is only hummingbird is such a bird that it can perform f(fly) as f(←(flying backwards)), adding to f(→(forwards))"]. Therefore, formally speaking, an essentialist definition of entity *e* is a 2-tuple e = <P {Ql, Qn, F}, F(x)>, where *P* is a set of {Qualitative, Quantitative, Formal} attributes and *F(x)* is a set of value-fixed functions.

3.3.2.2. Essentialist Approach Misused

However, when attempting to define intelligence from dispositions of essentialism, there are indeed: (1) improper applications of approach with regard to the subject (i.e., "nature of entity") to be defined; and (2) ill-formed operational frameworks. (1) here means that: what is meant to define intelligence is effectively unbound from its particular hosts, their essential feature and/or species-bound function realizations, as well as their constraints and limitations, physical constitution, history, or anything recapitulated within or merely given. Intelligence is a subject of theoretical and practical revisions conditioned by its open-endedness towards realizabilities, functions, means, and ends [where the latter are also often revised, repurposed, redefined, etc.].

With all that taken into consideration, the essentialist effort to define/outline what general intelligence (or AGI) is, would result into a cognitive failure: at the highest level of generalization of essentialist definitions of intelligence(s), a subsumption takes place of an epistemic "glitch" of *vicious inversion*—a subsumption of a *set* to its *element*. In our context, vicious inversion is defined as follows: A human-based, species-bound, or species-specific definition takes human species as a *paradigmatic model*, on which the definition of *general* intelligence is built, to different extents. Namely, *from* the denial of possibility of AGI realization due to substrate-bound implications—that is, relating to its artificiality as an obstacle, *to* the possibility of such a realization but—as it is given in a particular host of intelligence—in a form of essentialist replica of human-as-*quasi*-general intelligence.

The elements of an improperly justified transfer in essentialist definitions (as a subsuming principle that pretends to be self-sufficient towards the transferred content) include not only "functional invariances" of what it takes to be a host of intelligence, but also a contingent and unnecessary "base set" of species-related properties, species-specific [anthropocentric] bias, biological constraints, historically specific limitations bound to circumstances, or state-of-the-art proclaimed finality/totality of the givens, after which there is no historical unfolding assumed but rather an ossified "nonchalance." It also includes particular realizers that are treated as if they were function-definitive [:= value-fixed functions].

Reconsidering the example of the appropriate use of the essentialist approach, but in reverse, it is as if one tried to give a [general] definition of a bird of prey, enumerating, among others, a [species-specific] owl's vision, habitat of a species white-tailed hawk, and parametric features or functional properties that are observed in the Aegypiinae subfamily (Old World Vultures, one of the two subfamilies of Accipitridae)—and from this, the general notion of bird of prey should allegedly follow. A similarly improper equalization of scope, erroneous subsumptions, and violation of specification ordering/nesting principles are observed, although not reflected in a proper form, in essentialist definitions of AGI, general intelligence, *and* that of human *as a host of intelligence*.

Since the very concept of *human as a host of intelligence* here is neglected, rejected, or simply ignored, an ill-formed succession of what may be called (following Sellars) *images* of intelligence—not even *holistic* images but rather of arbitrary compounds—comprises the whole definition. Hardly ever is such a "chimeric" synthesis acceptable or usable at all. As a cognitive metaphor for clarification of what a mixture of "Sellarsian images" of entity may be about, consider a theory of mind comprised of assumptions, claims, postulates, hypotheses, descriptions, definitions, and representations such that one part of them deals with the ramifications of *folk-psychology* (common sense); another part is paradigmatic of its opposite, eliminative materialism; and the other part is simply copy-pasted from contemporary conventional points from cognitive neuroscience, adding all that to one psychoanalysis framing.

A thing to note: Not that *all* the claims at all scales here are false, wrong, imprecise, or vague. Some are true, correct, precise, and definite, but *this alone* is insufficient to turn the mixture into a consistent and adequate theory of mind. By the way… have you noticed a metacognitive bias in this metaphor? Right, this seems to refer to the *human* mind, not a mind in general. Which means that, my cognitive metaphor might not be as good as it may have seemed. Nevertheless, what this deliberate misrepresentation of *general* theory of mind actually shows is the ease with which one dismisses the anthropomorphic implications as inherent and, hence, unquestioned.

Defining correlationism, Quentin Meillassoux introduced the concept of the "correlationist circle" (Meillassoux 2008). This can be formulated as, "When you posit *X*, you *posit* X," referring to a nexus between *positing* and the *posited* (*thought of* X and *X* as external, thought-independent entities). Therefore, what (1) and (2) actually contribute is a viewpoint to an even tighter circle of correlation, reminiscent of a methodological, epistemological, and ontological "collar": With an anthropocentric conception of intelligence and AGI-correlationism, the case of a completely and abruptly ungrounded short-circuiting of the "AGI-correlationist circle" can be expressed as, "If one considers *intelligence*, one considers *anthropomorphic* intelligence."

Intermediate Conclusion

As observed from this panoramic review, the metaphysical, epistemological and ontological implications of correlationism are relatively strong within the components of artificial intelligence—both in theoretical considerations and potential future applications of philosophical attitudes to AI (such as AGI development). To address the problem explicitly, encapsulating the anthropomorphic and anthropocentric attitudes within AI discourse, the concept "AGI-correlationism" is introduced, as a use case for the broader concept "correlationism." It is also argued that the essentialist framework fails to define and otherwise grasp intelligence due to it being connected to species-specific [anthropomorphic] traits (including the speculative, not-yet-implemented AGI) in all notable and decisive aspects.

The functionalist paradigm, on the contrary, attempts to break intelligence down to what it does—in potential and actual registers of functions and capabilities. As asserted, the use of functionalist framework here is not only more consistent, robust, and comprehensive, but it also recognizes the diversity within the "sapience continuum," thus it can be aligned with the practical matters of AI development, favoring prospects of intelligence unconfined by human-based traits that are open to an unbound spectrum of implementations and realizations, resulting in an open-endedness aimed to adhere to the wholeness of spectrum for AGI realizabilities. The detachment from the anthropocentric view in terms of species and the quest for essential features of intelligence labeled as, "What it is," is not just a speculative exercise, but a necessary condition of progressive research and development in the AI domain.

References

Brassier, Ray. "Correlation, Speculation, and the Modal Kant-Sellars Thesis." *The Legacy of Kant in Sellars and Meillassoux*, edited By Fabio Gironi, 67-84. Routledge, 2017.

Hutter, Marcus. "One Decade of Universal Artificial Intelligence." In *Theoretical Foundations of Artificial General Intelligence*, edited by Pei Wang, 67–88. Atlantic Press, 2012.

Kaluža, Jernej. "Hume's Empiricism versus Kant's Critical Philosophy (in the Times of Artificial Intelligence and the Attention Economy)." *Információs Társadalom* XXIII, no. 2 (2023): 67–82. https://dx.doi.org/10.22503/inftars.XXIII.2023.2.4

Meillassoux, Quentin. *After Finitude: An Essay on the Necessity of Contingency*. Continuum, 2008.

Paksi, Daniel. "The coherent emergentist concept of machines; or why the popular concept of artificial intelligence is a materialist anthropomorphism." *Információs Társadalom* XXIV, no. 2 (2024): 85–100. https://dx.doi.org/10.22503/inftars.XXIV.2024.2.5