MSTYSLAV KAZAKOV

AGI-Correlationism and Its Discontents: Part 2.

This part of the paper systematically unpacks the most notable and decisive entailments of the implications of what was defined in the previous part by the concept of AGI-correlationism at all scales and levels, from the assumption that AGI must replicate human intelligence to the often-unquestioned idea of human-centric tests like the Turing Test. Representation of the entailments is then followed by their critical observation and discussion from the viewpoint of relevance, validity, truth or falseness, usability, and so on, arguing for another attitude, approach, and paradigm in all relevant domains (that is, the domains of reference of the entailments). The paper closes by some open questions that both parts of the paper leave at the end, and a closure.

Keywords: *artificial general intelligence, philosophy of AI, correlationism, philosophy of intelligence*

Author Information

Mstyslav Kazakov, National Technical University of Ukraine “KPI named after Igor 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 2.”.
Információs Társadalom XXIV, no. 4 (2024): 79–91.
https://dx.doi.org/10.22503/inftars.XXIV.2024.4.5

*All materials
published in this journal are licenced
as CC-by-nc-nd 4.0*

Following my first paper that dealt with establishing a theoretical framework of defining intelligence and an introduction to the concept of AGI-correlationism, the entailments of AGI-correlationism are now to be critically examined. Additionally, the auxiliary objective of the critical reflection is to preserve all the meaningful results of the first part of the paper, namely the concepts that were deployed so far, the pro-functionalist paradigm of understanding the A(G)I and its further usability, and to amplify these results by expanding the considerations. Hence, with regard to the latter, the critical part here struggles not to be merely 'destructive' or a deconstruction of the entailments of AGI-correlationism but also to propose some 'constructive'—that is, positive content—declarative theses and propositional knowledge (which, I hope, would be considered justified enough to be considered such, not merely pretending to possess such a status).

1. AGI-Correlationism and Its Entailments

Having done the theoretical framework, an observation and investigation should follow within this framework, of *what* correlationist attitudes (three questions from Part 1 of the paper: in section 2.1., further expansion of section 2.2., as well as instances of misuse of concepts and the approach itself, as is told of in subsection 3.1.2.2.) entail, and *how can* these entailments be fathomed in a comprehensive manner, and, that is to say, assessed according to their depths and postulates (how they are handled can be approached and reproached). It is a matter of fact that building here an exhaustive, all-encompassing enumeration of these entailments is an unfathomable task. I chose only a part, namely those which I consider important for AI ethics in general (thus, they cannot be ignored) and those which are the most ostensive for understanding AGI-correlationism. No particular tailoring point for these entailments to stem from exists. They are drawn from multifaceted sources, persons, dispositions, schools of thought in philosophy or AI research, of different backgrounds and causes of being as they are presented. To my mind, it is vital to specify and iterate the following entailments:

(1.1) The assumption that any AGI 'in a full sense' must be the replica of human intelligence concludes that intelligence of humans *taken as a species*, in its current form, is an *unsurpassable limit*. In other words, human intelligence in its current manifestation is considered as an *ahistorical* manifestation, a timeless measure and frame of reference to all things, and intelligence in particular. Such an implication is derivable not only from the idea of realization of AGI as brain emulation only but also from ethical commitments to abstract and concrete forms of anthropocentrism that imply a thesis on the uniqueness of human intelligence or just posit the idea of its being the upper limit of intelligence development (admitting the facticity of other intelligences in the history of the Earth and within the current state of affairs).

(1.2) *AI defined by pass-for-human tests*. Since the beginning of AI research, an anthropocentric attitude is explicitly exemplified by the 'imitation game,' later known as the Turing Test. In the best-known version of the test, humans engage in conversation with two or more hidden interlocutors, one of which is a computer (others being humans). If the interrogator fails to guess who is who, then the computer is said to be 'intelligent.' The very framework, even as a thought experiment without contemporary chatbots and other implementations, as it is presented in Alan Turing's *Computing Machinery and Intelligence*, presupposes completely anthropocentric expectations and attitudes toward the nature of 'intelligent computer.' This also may be extrapolated to any case where AI has to pass a test to be qualified as intelligent—*to pass as human*.

The same logic unfolds in some state-of-the-art (narrow) AI systems, known as '*artificial stupidity*'—top-down implemented suppression or constraint of AI performativity, algorithmically dumbing it down to such instances as deliberate erroneous outputs, due to a poorly or insufficiently implemented decision-making procedure. Another widespread tendency is making a system 'more convenient' in terms of interaction with the user, such that the latter perceives its features as more 'natural.' To make narrow AI look more 'natural' would mean *here* to look similar to '*superior'* general intelligence, human intelligence; particularly, in functional terms this necessitates the '*inferior'* intelligence (AI) to be prone to *making errors* instead of better responses for which it is capable, just because the '*superior'* human intelligence is susceptible to such mistakes. And all that is made for the purpose of AI passing the test on *pretending* to be human.

(1.3) *Intolerance to contingencies*. Consider a hypothetical future condition where the final draft of AGI is not fully equivalent to human in some domains of vital concern (such as ethical module, 'goals–means–drives' ratio, reasoning transparency, robustness testing, etc.). It is then going to be realized not as a human correlate but would have some contingent outcomes leading to unexpected and unpredicted (although not necessarily dangerous) emergent properties. AGI-correlationism considers such an implementation to be *unacceptable* until the contingencies are 'rectified' to humanlike or the predictability of the outcomes reaches some desired or 'sufficient' threshold, and emergent properties are prevented from development.

(1.4) Human-level constraints. Following (1), given the disjunction of the final drafts with three disjuncts at least, in other words, a possibility of realization of three different AGIs where one would be equivalent, by its realization *and* potential of development, to human; the second would be equivalent to human but with *development potential* of exponential growth in speed and far surpassing human level; the third one as already superior to human from the very moment of its realization (with contingent potentialities of unpredictable exponential growth). Given these three hypothetical final drafts, an AGI-correlationist would always choose the first disjunct, regardless of what the two latter may be (whether friendly or not, constrained ethically or in any other way, etc.), and how we can potentially benefit from them.

(1.5) What is also (implicitly) entailed by this attitude is an idea of superintelligence conceived in terms of *quantitative* more than *qualitative* superiority toward humans, which is, in a nutshell, reminiscent of (or analogous to) the distinction between humans and Olympic gods made in Greek mythology, where the superiority of the gods is actually measured rather in quantities than in qualities: physical strength, anticipation, skills in craft, terms of life (Olympic gods are not immortal), etc. The AGI-correlationist attitude toward superintelligence similarly reduces the superiority of the latter to all-too-human (as-species) traits, areas, and matters of concern, thus imaging superintelligence as (hypothetically) being merely a human individual with unfathomable strength and intellect. And just as any human, due to its 'corrupted nature,' it would necessarily get drunk on its own strength and would start seeking absolute power (world-domination and beyond). So, if there would be a potential for AGI to become superintelligence, whatever the premises or presuppositions, to an AGI-correlationist, it would be necessarily conceived as a malevolent human with unlimited intelligence. In a nutshell, as one of the potential consequences, the very possibility of realization of AGI with a potential of becoming superintelligent, would possibly be rendered into one more premise of abandoning of realization of AGI (as an addition to (3), (4) or both).

(1.6) AGI as intelligence-*for-us*. AGI-correlationism postulates that the AGI that we realize, whatever its factual realization may be (the level of intelligence, possession, or absence of self-perception as an individual, etc.) must, regardless of the implementation or the nature it exhibits, deeply care about humanity in general, and us specifically, focusing all the knowledge, aspirations, and desires bound explicitly to humanity—human causes, concerns, interests, goals, problems, issues, inquiries always to come first, that is, prior to those of AGI itself, whether the latter possesses any of the aforementioned. The explicit example of such an attitude may be *preference utilitarianism* as it is laid out by Russell (2019), based on what is known as the 'first principle of beneficial AI,' which may be formulated as follows: the only purpose and objective of machine AGI is the realization of *human preferences*. At the very core of the conceptions akin to 'beneficial AI' lies the implication of AGI as 'for-us' in the described manner, which, from the viewpoint of ethics, is in fact a nullification of the recognition of AGI as an intelligent entity (if to follow the previous idea of mutual recognition as a foundation of relations between human and AGI).

(1.7) AGI-as-(public/private)-property. This entailment is similar but not identical to the previous one. For some thinkers, including prominent AI-theorists (Coeckelbergh 2020; Bryson 2010), possessing 'full-scale intelligence,' including 'analogous to humans' or surpassing it (in every domain of interest or at least in some of those) does not necessitate a change in treating AGI in a way other than 'machinic property,' similar to a bicycle, a fridge, a TV, or a laptop: whatever its level of intelligence is, it is still reified, treated not as an individual or a host of intelligence in general, with corresponding 'obligations' and attitudes toward it from the perspective of humans. It is argued that "the status of AIs will be ascribed by human beings and will depend on how they will be embedded in our social life, in language, and in human culture" (Coeckelbergh 2020, 59); but if *this* is the case, then any realization of AGI, whatever its properties, the more 'uncanny' they are (divergent from humans), the less equally they would be treated.

Not only does the problem lie in explicitly uncanny features: to the register of 'uncanniness' of AGI may also be related the failure of 'desired' supplementary (secondary) properties, such as, for instance, explainability and transparency (hypothetic AGI properties referring to the possibility of either self-explication of the results or actions of AGI by itself or the backtracking of such results, actions, and their underlying causes (code, algorithm, trigger, intermediate steps etc.) by humans. Artificial neural networks are *already* conceived as black boxes, precisely because the number of layers of artificial neurons already involve computations so intricate that satisfiable human affordance of backtracking the outputs vary from extremely hard to impossible. A similar concern is expressed in one of the latest publications of OpenAI researchers:

> If a superhuman assistant model generates a million lines of extremely complicated code, humans will not be able to provide reliable supervision for key alignment-relevant tasks, including: whether the code follows the user's intentions, whether the assistant model answers questions about the code honestly, whether the code is safe or dangerous to execute, and so on. (Burns et al. 2023, 1)

Yet, generally speaking, 'uncanniness' not only problematizes further advances in artificial neural network design or a problem of supervision for what is known in AI research as 'superhuman models'—if, in a hypothetical future, a hypothetical AGI would not fill in this gap by itself, or if no solution exists for the models mentioned given that "naively using weak, human-level supervision will be insufficient to align strong, superhuman models; we will need qualitatively new techniques to solve superalignment" (Burns et al. 2023, 8)—it would remain an opaque and unexplainable 'black box,' contributing to its uncanniness (as conceived by humans), such that it may invoke a more negative, prejudiced, and biased attitude toward AGI from the viewpoint of humans, as it could have been in case of explainability and transparency realized as its properties to a certain extent.

2. Entailments of AGI-Correlationism Critically Observed

Having these points explicated from discursive 'hum' and enumerated, in this part I will now attempt to focus on their critical exploration and decomposition, on the one hand, trying to explain their nature (which itself is, to my mind, the most effective and unmatched form of philosophical critique if carried out consistently, grasping the nature in its explanatory model and representation). On the other hand, here I am also trying to give an outline of their 'what is wrong fundamentals' of these entailments, their attitudes and dispositions. This observation is not exhaustive or all-encompassing (no observation is), and, given this, it also serves as an open-ended invitation for discussion.

(2.1) One of the underpinnings of AGI-correlationism is a metaphysical/ontological stance commonly known as *exceptionalism*—an implication of uniqueness, being 'one of a kind,' here referring to an attitude toward human intelligence as synchronically and diachronically *unique* by its functions, features, and capabilities in orthogonal characterization. Such are the essentialist/correlationist implications, which are, at the same time, consequences into which one arrives basing on such implications. From the viewpoint and general dispositions of functionalism *and* historicism, human intelligence is not unique, not merely speculatively (because of the possibility of realizability of AGI) but also empirically; it is regarded as already proven to be functionally 'ordinary' or common—by archaeology.

The first, wider renown, is the discovery of a site of Oldowan toolmakers in 1934, which for a long time was considered to be the oldest stone tool industrial site ever known—with the corresponding consequences. In addition to it, aside from either a late Australopithecus or early Homo habilis, our closest phylogenetic cousin, one of whom assumedly created the complex, there is a more recent finding with more enigmatic outcomes. In 2011, a group of archaeologists led by Sonia Harmand, during excavations in Kenya at the archaeological site Lomekwi 3 found ~150 (20 well preserved) stone tools dated as being 3.3 million years old. The artifacts themselves in their variety, dimensions, and visible percussive-related traces on the artefacts led the archaeologists to suggest that the hominins that made them combined battering practices and core reduction, using the artifacts variously for agricultural purposes, as cores for flake production, for pounding, or as anvils.

Several distinctive uses of individual objects for multiple tasks presuppose goal-making capacities in the form of goal-representation, reflecting a degree of technological diversification higher than had been thought possible for the period, before the findings of Harmand's group. The assemblage may be an instance of a technological stage between a hypothetical succession from pounding-oriented stone tools used by early hominins and the flaking-oriented knapping strategies of Oldowan toolmakers. The artifacts at the Lomekwi III site are *at least* 700,000 years older than the tools at the 1934 find; they also predate arrival of the whole *Homo* genus, our phylogenetic family, by 500,000 years. The question of what hominin species made them is still open, save for their being relatively detached from our closest ancestors, with an intelligence level of a potential no less than our own. Questions about the cause of their extinction, their loss, and the reinvention of their technology (or we just have not found other sites showing the succession instead of reinvention) are open as well.

Speculative derivations from the discoveries may be as follows: *if* intentional, non-spontaneous, organized, consistent, and serialized practices of collective stone toolmaking is sufficient to consider the agent as having at least the same potential of intelligence as humans do, then humans are *at most* third in being such agents, since being predated at the very least by a distinct species that had the same intellectual capacities; and it belonged to *genus* (predating our own by at least 500,000 years). Following reflections of early Marx (in 1844 manuscripts) and expanding his thought, it is worth juxtaposing three different modalities of toolmaking. The first one is a genetically 'embedded' drive similar to that of woodpeckers or termites, beavers or spider monkeys, involving them in niche-constructive ecological engineering of various complexity and ecological significance. But this is not the case for Oldowan or Lomekwi III toolmakers, as anthropology, archaeology, and evolutionary biology conclude. The second modality is the 'top-down' automation of production we observe in machines and/or robots, when a human (or another machine) prefigures and programs a certain technological means of production to create a material artifact, which is obviously not the case either. The third modality to consider is teleological toolmaking—the one caused by the *goals* and *needs* of some sapient being, a decision-based outcome. And this latter *is* the case, an artificialized functional extension of mind and body of the organism, which may serve as a criterion for ascribing intelligence to an organism. Adherent to this reasoning, given the facticity of at least two empirical instances of this third modality of toolmaking in the prehuman history of the Earth, one is eligible to infer the strong claim asserting the denial of human intelligence exceptionalism. (Perhaps, in other contingent circumstances, it may have been that *they* had become *us now* instead *of us* as possessing 'unique intelligence' as some think we do. It may have been that both groups had gone extinct at approximately the same prehistoric time.)

(2.2) Another principal flaw inherent in any AGI-correlationist conceptions and views of top-down alignment or any other instance and degree of design concerning values, decision-making procedures, rule-governed behaviors as both constraints and space of choice between following one of the rules in case of disjunction where the rules to follow (actions to be taken are disjuncts), is the instant absence of a common axiological 'denominator.' Even if we come to some more or less common instances for abstract and ambiguous terms like happiness, justice, dignity, respect, autonomy, stable development, value of life, and so on, we will not be able to proceed with precision for each instance without arriving at a contradiction.

For example, one may try to take something 'basic' and 'obvious' from the list of prioritized points of alignment, regardless of the final draft of AGI, such as 'value of life', and attempt to perform a 'fine-tuning' of it *as* a factual subject of alignment, before setting it as a top-down 'directive' for an AGI. To start with, the state of the art in the scientific domain we generally refer to as 'life sciences' is that there is still no, as Eugene Thacker (2010) demonstrates, *positive* definition or a sufficiently posited concept of *life*. That is, life sciences can extensively enumerate *what life is not*, but they are still struggling with affirmative claims of *what it is*, generally, i.e., without reference to particular *livings* (as Aristotle has it), or *life-forms* (as we usually phrase it today). Can, then, alignment concerning it be laid out as 'what is *not X, not Y, not … ad inf.* must necessarily be valued and, in any instance, prioritized', or 'another matter/subject of care/concern/valuation should be the following array of entities: extremophiles, bacteria, …., human, whale, trees, …., *the whole 'tree of life' is enumerated'? Obviously, it cannot. But suppose the future science at the dawn of AGI would effectively resolve this ultimate scale challenge, letting us effectively move down to further details of value-of-life alignment. What about ethical conundrums and clashes between: pro-choice and pro-life proponents? Vegetarians/vegans and omnivore diet advocates (and the gradations between the two poles)? Pro-/anti-death penalty? The list of contradicting, mutually exclusive, poorly or absolutely incommensurable dispositions directly related to the 'value of life' is infinite.

These musings so far are not original, and the overall state of affairs here has not changed since Bostrom's *Superintelligence* (2014). But here is the unexpected entailment. Of course, a more advanced intelligence is potentially *capable* of convincingly and consistently resolving these conundrums, either by conducting research to which we are not intelligent enough to do or just by reasoning to an unequivocal conclusion, effectively verifying or refuting a particular axiological disposition. To solve the abovementioned ‘unresolvable questions,’ antinomies of (human) mind is actually one of the most desired potential benefits expected from AGI! *And hence*, to have a sound and ‘proper’ alignment with, say, ‘value of life’ or any ambiguous concept of the domain, *we must let AGI decide for us* what *this value is,* and *how* it should be conceived, what *is* relevant to it, and so on; and eventually it is *we who* are the ones to be *aligned* with *AGI* and its judgment, value system of coordinates, and measurement scale!

Ill-defined or ambiguous goals and priorities of/for humans having, as one of the consequences related to the subject under discussion are a well-known problem within the whole domain of AI ethics, and its AGI subfield, which remains as speculative, since the perspective of AGI realizability is remote at the current moment. In addition, there is the more instant, more immediate, and more technical issue of ‘superalignment’: given a model with superhuman abilities (not necessarily AGI, but also a narrow model that is exponentially superior to humans in some particular areas of vital importance), how should it be supervised and aligned so as not to perform erroneously having catastrophic consequences? More briefly, how can weak teachers/supervisors teach/control a model that is much smarter than they are?

Although some solutions, such as weak-to-strong generalization, have already been proposed, it is an open question whether the methods would be able to keep up with the pace of the growth of intellectuality (not speaking of an “explosion”). Even if they could, superalignment would always remain a form of *mediated* alignment where humans would instantly rely on other tools *as* methods of alignment of supermodels, unless they become superintelligences themselves. I bring this issue out here to imply that, even concerning the ‘all-too-human’ (seemingly) matter of alignment and supervising the AI models, there are already matters of concern that are forever out of the correlationist implications, attitudes, and scopes.

(2.3) The problem with anthropomorphizing tests is not only the way they instantaneously ignore human biases (e.g., incompetence in evaluation and diagnoses) but also (and, perhaps, *mostly*) in the general reasoning which underpins the conception of such tests and attempts to justify them (preserving their use even today), such as: ‘Yes, biases, yet *we have no other options*’; ‘The concept of intelligence must be defined first’; ‘No one would take the results seriously without a pinch of salt!’; ‘What is wrong with humanlike interfaces for the applications *made for humans*? Would you prefer a car with a traditional interior or one without a seat and with pedals above your head?’; ‘Human nature itself is not fully explored and cognized; therefore, when AI tries to pass for human, as you say, the very fact of nonhuman *X* passing for human also gives us additional clues about ourselves, of what it is ‘to be’ human.’

Such arguments are actually not completely irrelevant. However, when broken down, they also hold that the very idea, cognitive metaphor, fact or even image of a genuinely intelligent (sapient) but inhuman entity is something intolerable or even *immoral.* In this respect, the above claims are unjust as emotionally laden and, thus, being more value judgments than arguments. There are objections even if we take them for ‘full-blooded,’ valid arguments. More or less general, precisely addressed toward the whole idea of ‘tests to pass as a human,’ that I consider as one of the most convincing, since it is affirmative and constructive instead of being merely critically 'destructive,' comes from the viewpoint of a formal approach to AI research known as universal artificial intelligence, "a new paradigm to AGI via a path from universal induction to prediction to decision to action" (Hutter 2012, 69). Emphasizing the obsoleteness and inefficiency of anthropocentric tests, Marcus Hutter instead calls for non-anthropocentric tests of and for intelligence, which emerged in the last decade, such as the universal C-test inspired by Solomonoff induction and Kolmogorov complexity. This test, as well as the others of its kind, are centered around task solving and learning-from-scratch, for a purely detached 'agent–environment' interaction without any of the two alignments with human traits (at least as much as such a detachment is possible).

(2.4) Just as with 'paleohumanism', antihumanism, misanthropy, or extropianism, AGI-correlationism is *intolerant to indifference*. As Benjamin Bratton puts it, "[p] erhaps the real nightmare, even worse than the one in which the Big Machine wants to kill you, is the one in which it sees you as irrelevant, or not even as a discrete thing to know. Worse than being seen as an enemy is not being seen at all. Perhaps it is that what we really fear about AI" (Bratton 2015, 70). Surely, this mustn't be extrapolated onto any instance of correlationist attitude toward A(G)I, as well as non-correlationist attitudes—as something that we *all really* fear, for it is definitely not the *greatest* matter of concern. At the same time, fear of indifference is *among* such matters—just like this fear/intolerance underlies general metaphysical/ontological attitudes toward the indifference of cosmos to mind, as it follows from the great humiliations of human; as Meillassoux argues, philosophical 'disagreements' with cosmic indifference—found in Fichte, Heidegger, Derrida, Husserl, Kant and others—are the 'firestarter' of correlationism. It is no surprise that rejection of indifference is among the expectations of humans toward AGI, since it is prefiguratively anthropomorphized.

(2.5) Another important concept/problem that AGI-correlationism fails to grasp properly, i.e., without anthropomorphizing it, is *agency*. In the discussed context, a threefold distinction should be conceived properly. Firstly, we have the property/ feature of *agency*, which itself is definable in a broader context, as we have it in philosophies of the mind, theories of action, ethics, or cognitive sciences. Secondly, there is a specific, 'narrow,' technical use and meaning of the concept '(intelligent) agent' in AI research and practice divorced from the use of this concept in a broader philosophical context or common sense implications. If philosophers (or any other non-AI theorists/developers/deployers) would like to discuss the entities to which the concept 'intelligent agent' refers, they (necessarily) adopt the 'narrow meaning,' which *is not included in a broader one*.

Thirdly, there is an even narrower technical term, divorced from both of the two above, referring to the property of *some* AI systems—hypothetical *and* already existing—known as *agentic* AI systems, which are not identical with intelligent agents. Agentic AI systems are distinguished by their ability "to take actions which consistently contribute toward achieving goals over an extended period of time, without their behavior having been specified in advance" (Shavit et al. 2023, 4). The property attributed to them is encapsulated in the concept of *agenticness*, referring to the degree to which an AI system is capable of adaptably achieving complex goals in a complex environment with (relatively) limited direct supervision and a wide independent execution permission (the degree of the system's autonomy of actions and decisions being considered reliable without the need for approval from a human user).

AlphaGo is an intelligent agent, but it is not an *agentic* system in this described sense: it is incapable of complex goal accomplishment, does not have independent execution permission aside from playing the game Go, and its environment is static (and not augmentable). Google DeepMind's new *FunSearch* is, on the contrary, an example of an agentic system, being *highly* adaptable to complex environments, comprising a systematic evaluator paired with *Codey*, a pretrained large linguistic model that is itself a version of Google's PalM-2, fine-tuned on computer code (the principle of work and goal-receiving is made on the same basis—e.g., a mathematical task in set theory input in Python with a response/solution in a form of executable code(s), the most prominent variations of which are selected by verifiers). Its direct supervision is notably limited; it solved the cap set problem, a mathematical problem of extremal combinatorics that remained unresolved by humans until the end of this past year, but, as *FunSearch* creators admit, they cannot fully understand *how* this discovery was made, since the whole way cannot be backtracked. At the same time, its independent execution level is more limited than that of agentic AI systems implemented in cars capable of level 3 autonomous driving. (*FunSearch* provides the user with a code that *can be* executed, although the decision whether to execute in or not, *which of* them, given multiple variants, is ultimately up to the user, while in level-3 autonomous driving cars the crucial executive decisions in critical situations are made independently of humans.)

The provided example sheds light on the more subtle and abstract 'philosophical' nature of the difference between *agency* and *agenticness*. While the former has classical 'Boolean' truth values (either it is *present/observable* or *absent* in an entity under valuation), the latter has 'fuzzy' or 'many-valued' truth values. The ultimate value is a matter of degree, *from* a minimal threshold to a non-specified upper value, determined either by contrast and comparison, or on the basis of evaluation by criteria, the list of which may be subject to change and a proposal from potentially different research programs. In Shavit et al. (2023), four such criteria are proposed: goal complexity, environmental complexity, degree of adaptability, and degree of independent execution or action-space autonomy. To simplify, in the case of agency, either X *is* an agent or *it is not*; while, given n ≥ 2 agents, an arbitrary X may *possess more or less* agenticness than any other agent Y. It is also worth generalizing that, although it *correlates* with *generality of functions/capabilities* and *task-performativity* of an AI system, there is no *necessary nexus* between these properties (for instance, many digital systems are more agentic than almost any AI embodied in a physical robot). And if AGI has any perspectives, agenticness is as important a feature to consider as generalization across domains.

Now, back to the point. It is crucial that agenticness must be perceived distinctly from such things as: mind, consciousness, moral agency, awareness of self/others, motifs and motivation, etc. From this follows that its meaning, content, and value are independent and distinguished from the (actual or given) degree of a system's *anthropomorphism*. Agenticness neither implies nor requires "a human-like appearance or human-like behavior, though anthropomorphic appearances and behavior may increase the likelihood of humans perceiving such systems as agentic" (Shavit et al. 2023, 5). For AGI-correlationism, *it is* the system's anthropomorphism which defines, and *what* it defines is *agency*, hence, the latter is not 'Boolean,' but 'fuzzy', and what in turn determines a degree of agency here is a degree of *autonomy*, in a general sense. The degree of autonomy is conceived, for instance, as the qualitative and/or quantitative limits of freedom given to a particular subject of decision, choice, and actions, determined by external conditions and circumstances (instant and contingent, correspondingly) and the subject's internal nature.

In this particular case, not everything should be reduced to and explained by the issue of anthropomorphism. While not being *false* or 'generally' wrong, the problem with such a representation of the state of affairs is *obsolete* with the state of the art in AI R&D. But that is a general disposition of correlationism—undesirability of resignation of changes and generally what we may call 'revision' or 'renegotiation'—of concepts and definitions, frameworks, theories, strategies, approaches. This is not because correlationism is opposed to the ideas of progress, evolution, adaptation, and all similar things; this implicit denial of changes stems from essentialism, which tries to ossify any given state of the art as *totality* or *closure* by absolutization of reified successive stages as a final/ultimate singularity. This, actually, is what makes correlationism a bad counterpart to AI ethics, precisely because the rapid and contingent path and pace of development necessitating renegotiation and introduction of the new concepts, methodologies, and definitions *is what* makes up this development. To put it short, AGI-correlationism, underpinned by essentialism, tends to retain some generalized points, as well as particularized issues that are obsolete or redundant, resulting in theoretical and practical misrepresentation of the state of the art in contemporary AI. Anthropomorphism as a criterion for *agency* evaluation *used to be* 'wed' to what is defined here as *agenticness*, referring to the 'GOFAI dispositions' rather than anything meaningful today.

3. Open Questions

A lot of questions are left unanswered here and some have not even been tackled. Therefore, whatever closure follows, the subject matter remains open-ended. With this abductive principal open-endedness in mind, this section addresses some of these questions, taking premises for future investigations, accessible not only from the 'outside,' but also being self-critical, an attempt to see one's own weak points, calling for improvements and future work.

Q1. What are potential or actual perspectives of use of the introduced concepts and arguments in further philosophical investigations? What *affirmative* and *constructive* criticism can/should be added to part 2?

Q2. How should we actually demarcate between what was called here 'human-centred AI ethics' and 'AI ethics of correlationism'? Should there be a checklist, threshold, etc., by which one may effectively distinguish between the two?

Q3. Should some points and arguments of correlationism concerning specific topics, as in the case of (3.3), be accepted and reworked, and integrated into another AI ethical system? And what about the contrary? Aren't particular concepts introduced in the paper, along with others taken as they are, in need of renegotiation? This is because some of them are vague and poorly explicated (e.g., 'recognition').

Q4. Functionalism is an *approach* for theoretical frameworks and research itself. That is totally fine, but how about ideas about the *methods* of research? Should they be 'purely philosophical' in kernel, technical, multidisciplinary, and, if the latter is the case, would they be given any precision in order not to remain 'formal' proclamations of multidisciplinarity?

Q5. Given the state-of-the-art in our fathoming of AGI (not even knowing the path which leads to it), is it not too early for academic discussion of this kind? Is there any comprehensible evidence or argument that would ensure that the speculations of this kind would not be dismissed as 'transhumanist values,' wretched to be parochial precisely due to almost-imminent binding of them in 'at-hand' history, politics, and society?

4. Conclusion

At the present level of the results in research and development of AI, an area of epistemic enquiry itself one of the 'youngest' so far and done 'from scratch,' unlike many other subjects, having no preceding debates or rooting in past centuries, we are still far away from the creation of AGI: not taking into account any contingent or unpredicted breakthroughs in the field, it is not even 'the dawn' of it, not speaking of its potential 'rise,' or emergence. All this makes such discussions highly speculative. Perhaps it is the only subarea of general ethics where the latter is a matter of speculation more than a practical affair. At the same time, just as with a field of security in the AI domain, it belongs to questions where speculation *should* come first, because, when there would be conditions of actual realization of AGI, we must be prepared for the mitigation of risks it potentially brings along with the potential benefits (security), as well as having strategies of long-term interaction with it, such that both species would benefit from this interaction without doing harm to one of the subjects of the interaction or both (ethics).

To effectively deal with the negative consequences of anthropocentrism, anthropomorphism, bias, and the problems they lead to in the field of AI ethics, philosophy in general, as well as possible harm to AI research and related domains of thought and practice, not only critique of the attitudes, but also a framework of understanding and representation, unifying critique, explanations, definitions, and constructive claims (alternatives, propositions, reviews, and renegotiations). My intention was to demonstrate how a specific general approach (to definition, methodology, paradigm, etc.) can be impaired with particular ethics, and what such an impairment may yield, for good, for neutral, and for bad.

Essentialism and correlationism reciprocally imply each other. Functionalism implies different realizabilities, from a fully sovereign AGI agent to human-centred ramifications. The latter consider recognition of AGI as an autonomous and intelligent entity, with interactional parity. Considering human in the first turn as a host of intelligence, taken as a positive object of concern, instead of human as a species, it is possible to build up a common cause, a necessary condition for aligned, constructive, and efficient interaction between human and AGI. To mitigate ethical tensions, complications, and other negative outcomes, we should consider the possible causes, and I propose AGI-correlationism as a framework in which such a mitigation is possible to be performed.

References

Bratton, Benjamin H. “Outing Artificial Intelligence: Reckoning with Turing Tests.” In *Alleys of Your Mind: Augmented Intelligence and Its Traumas*, edited by Matteo Pasquinelli, 70–83. Meson Press, 2015.

Bryson, Joanna. “Robots Should Be Slaves.” In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*, edited by Yorick Wilks, 63–74. Amsterdam: John Benjamins, 2010.

Coeckelbergh, Mark. *AI Ethics*. The MIT Press, 2020.

Harmand, Sonia; Lewis, Jason E.; Feibel, Craig S.; Lepre, Christopher J.; Prat, Sandrine; Lenoble, Arnaud; Boës, Xavier; Quinn, Rhonda L.; Brenet, Michel; Arroyo, Adrian; Taylor, Nicholas; Clément, Sophie; Daver, Guillaume; Brugal, Jean-Philip; Leakey, Lousie; Mortlock, Richard A.; Wright, James D.; Lokorodi, Sammy; Kirwa, Christopher; Kent, Dennis V.; Roche, Hélène. “3.3-million-year-old stone tools from Lomekwi 3, West Turkana, Kenya”. *Nature* 521, no. 7552 (May 2015): 310–315.
https://doi.org/10.1038/nature14464

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

OpenAI. Burns, Collin; Izmailov, Pavel; Kichner, Jan Hendrik; Baker, Bower; Gao, Leo; Aschenbrenner, Leopold; Chen, Yining; Ecoffet, Adrien; Joglekar, Manas; Leike, Jan; Sutskever, Ilya; We, Jeff. “Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision.” Accessed December 27, 2023.
https://cdn.openai.com/papers/weak-to-strong-generalization.pdf

OpenAI. Shavit, Yonadav; Agarwal, Sandhini; Brundage, Miles; Adler, Steven; O’Keefe, Cullen; Campbell, Rosie; Lee, Teddy; Mishkin, Pamela; Eloundou, Tyna; Hickey, Alan; Slama, Katerina; Ahmad, Lama; McMillan, Paul; Beutel, Alex; Passos, Alexandre; Robinson, David G. “Practices for Governing Agentic AI Systems.” Accessed December 27, 2023.
https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf

Russell, Stuart. *Human Compatible. Artificial Intelligence and the Problem of Control*. Viking, 2019.

Thacker, Eugene. *After Life*. The University of Chicago Press, 2010.