Commentary

Artificial Intelligence Like Humans; Humans Like AI: Epistemology of Analogy and Our Expectations Beyond It

Ana Bazac[1]

1. Division of Logic, Methodology and Philosophy of Science, Romanian Committee of History and Philosophy of Science and technology, Romanian Academy, Bucuresti, Romania

In this paper – which has in its background a semi-joking smile – I propose an optimistic image of Artificial Intelligence (AI) considered in its plausible inherent development and *future* as a *new cognitive entity, that is, a new thinking entity*. This proposed thesis is the result of an *epistemological* approach that emphasises the common/shared role of *analogy* in both human cognition and AI's inferential response to its environment. In turn, the stages of analogies in physics highlight the contradictory beingness of AI, but this contradictory beingness is not specific only to AI, even though that of humans is of a different nature. Anyway, AI's efficiency is precisely the result of its larger field of data and information for analogy, and thus of its *much better* answers to the problems of the world. But could this larger field not also be the basis of better human knowledge and values as reasons-to-be for actions? Of course, the scope of judgements reflects "the input", information as the object on which they are exercised. Accordingly, and conversely to the present banal approach of AI as a copy of the human, AI can be a model for the treatment of humans by humans. So, as in billiards, in this paper the focus on the epistemic features and role of analogy in cognition is only a way to support the meanings of human access to information. However, if the critical spirit, as a result of the free access to information for all humans, highlights the problem of what marvellous things they can do on this basis, the development of AI on the foundation of humans' free analogy opens questions related to its existence alongside its creators.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

 doi.org/10.32388/N8JU5J

Table of Contents

1. Introduction

1. This paper in the area of *epistemology* only suggests the inevitable[1] development path of AI[2] as a *new cognitive entity*, alongside its natural model, the human being; this is consonant not only with the dreams of the human engineers of AI but also with the entire technical process of its construction. So: AI is based on and makes *analogies related to the real world*, as this is the raison d'être – as constitutive trigger and *telos* – of every intelligence.

2. The constraints/limits of this paper consist of:

a) AI is considered only as its *model beingness*[3] – and not in its *present* guise[4] –, as a *model* of its *future cognitive* development, extrapolating the present capabilities and directions of epistemic "gain of function"[5] in up-to-date AI science; the assumed model beingness of AI excludes sentiments, feelings, and abilities (given in humans to a great extent by the body), in other words, it excludes the

communicated uniqueness of the general cognitive experience, aspects which are in fact *sine qua non* for and in *human* intelligence[6] that, obviously, is the *criterion* and the *model* of the AI being;

b) a declared narrow *epistemological* view of AI, concerning only the role of *analogy* in AI reasoning, similar to that of humans; therefore, the theoretical reduction of the (model of) future AI to the structural/constitutive cold means of cognition as an abstract epistemic process is *itself reduced to the relation with/towards the information base*, the key to efficient analogies and thus, to knowledge;

c) the implicated optimism that this AI model is/might be a window to humans' necessary *free access to information*, namely, *to a critical standpoint* concerning the social information and contexts within which it is generated and communicated.

3. The paper's subtext is not a joke but an allusion. Rather, the allusion itself determines us to question why it is made.

4. Obviously, first of all, *the kind of AI* people want is the problem: a possible equal *colleague* of humans[7], or a good new type of *tool* that unites "the causal reasoning abilities of our best scientists with the sheer compute power of modern digital computers"[8]. Both alternatives suppose a high technical endeavour and bet, but socially the second is cherished: not only by laymen who fear the loss of jobs and *subordination* to AI[9], but also by AI professionals[10]. Actually, the second alternative is sold to present consumers as a good or bad substitution for endeavour and creation, covering laziness and sufficiency[11]. Accordingly, the first alternative is the reverse and, more, challenges human creativity in a *kosmos*/order with (only?) two "rational beings" in the Kantian sense in which *reason inherently involves the universalistic moral of the categorical imperative*[12].

5. Further, generally, attitudes towards AI regard its *present* stage[13] and its immediate *use*[14], and stretch from an optimistic "*technophile*" view of its social use – an *ideology*, and not a neutral technological enthusiasm, promoted by the leading technological and political centres of world power[15] – to an inverse warning that, in its turn, reflects both a valid analytical preoccupation[16] but also, and not necessarily included in the analytical view, a historical, "context-dependent" *technophobe* ideology of technology as the "automatic" Cause/main cause of social stability and transformation. As an analytical attitude, it is about a *critical* view that mixes optimism and concern[17], at the same time sending to obviously important but *resulting* aspects of the capitalist structural relations. As an ideology, technophobia is as *a-* and *non-critical* as the official technophile discourse.

6. The present ideological attitudes towards AI – reduced to a cognitive device as an entity, or imagined as a complete being "because it knows" – consist either of its reduction to a subordinated and beneficial instrument *or* its substitutive power over humans. Paradoxically, for the time being, AI is used not only as a marvellous *necessary* tool in science and human civilisation but also, by those who control its *use* and consider it euphorically a subordinated device, as an exaggerated and unnecessary substitution for the beneficial strain of human efforts of understanding, imagination and creation[18], inducing a mental laziness and intellectual backwardness that fosters just the catastrophic view of "our substitution by AI": the official optimistic model of AI turns nightmarish.

7. My paper does not concern AI's use but its *epistemic formation through analogies*: and albeit this formation is analogous to that of humans, the model of a fabricated intelligence shows, optimistically, the necessary conditions of human intelligence.

This epistemic framing is, however, not a (voluntary) closing within an ivory tower. Rather, on the contrary: because the evidence of AI's epistemic formation is its use. And the use is regulated *through epistemic interventions*, so that the goals deployed by AI remain within the field of the acceptable[19]. The problems posed now by Grok (**31**, **32**) show these paradoxical relations between technical exploits and social brakes and horizons.

8. Both the feeding of AI with biased ("discriminated") data, thus determining the generation of biased answers – be they the official Unique Truth or the "conspiratory" "fake news" which, according to official positions, would institute the "post-truth" era –, and the ubiquity of AI, that is, the alienated guise of humans who lie under its power, are new phenomena only when considering the *AI mediation*: actually, modern people *were* fed with prejudices and lies long before AI, so these prejudices and lies existed before and exist also outside AI; and both the *public agenda* and the *public models* were and are systematically imposed long before AI and also outside it; *first and foremost*, alienation does not consist in the cancellation of privacy by AI – towards which there can be instituted and there are preventing rules[20] –, but in: a) the determination of the qualities and quantity of data, information and conditions received by people in the context of power relations, and b) their alignment and subordination to them.

9. The epistemic similarity of AI and human reasoning based on analogy is interesting and significant not as a basis of an imagined ideal AI – because this ideal projection is confirmed only by the future, which is open. But the epistemic similarity allows a *possible* model of AI, which is necessary and efficient here not so much for outlining paths of its own development but for the treatment of humans by humans. If the performances of AI depend on the data and information it is fed, the same condition

determines the performances of humans; and, because these performances are qualitatively and quantitatively more and higher as the human basin as such is larger, access to a free, large and diverse pool of data and information is absolutely necessary for the performances and capabilities of *human* survival of the *human species* as such.

10. Therefore, there are two reasons for optimism concerning AI: one is the common epistemic ground of humans and AI, and the other is the possible optimisation of humans' free access to information, thus to knowledge, thus to thinking, as the AI model sketched here shows. But the smile covers both the doubt and hope that reasonable scientific models could overwhelm the deep political interests of the domination-submission structure in our real world.

2. Understanding analogy: Kant

11. Kant drew attention to the principles of "pure understanding"[21], namely of the *mental pattern* of *recognition* and *adequate reaction* through the *processing* of mental data of entities (empirical notions and abstract concepts), properties and relations. "The principles" are and describe *preconditions* of understanding as such, thus of cognition that is "through concepts", not reduced to intuitions as "sensible intuitions...grounded on the receptivity of impressions" "but discursive", and the *concepts* being means of understanding.

Epistemologically, to understand is *to judge*, and *judgements* relate *representations* of objects on which the human focuses[22]. But, since only sensible intuitions represent the object directly, judging means *to relate* both the direct and indirect representations, the concepts[23]. And, once more, of course, cognition is more than judging in order to *react* – nowadays we speak about "access consciousness"[24] – it is "thinking", "cognition through concepts" as "predicates of judgements"[25], that is to say, an autonomous inner articulation of infinite relations in the frame of the infinite experience that generates data and information.

Nevertheless, the understanding of objects is the basis of cognition. Consequently, it must be understood as a mental *process*. The framework of this *process* is that of "the principles of the pure understanding", viz. of phylogenetically strengthened mental models functioning as *schemes*[26] or *forms* for the processing as such of representations into coherent units. These principles consist of *the axioms of intuition*, *the anticipations of perception*, and the *analogies of experience.* Thus, both i*ntuition*, *perception* and *analogy* explain how the unknown comes to be known. The first analogy is based on the learned frame-idea of

"persistence of substance" and fixes it[27]. The second analogy is based on and fixes the frame-idea of "temporal succession according to the law of causality"[28]. The third analogy is called the "Principle of simultaneity, according to the law of interaction, or community"[29]. All analogies are fixed frame-ideas without which the production of novel, not-yet-discovered ideas is not possible[30].

3. A word on methodological levels of analogy

12. There are some methodological levels of analogies/analogical cognition. The first, the *formal* level, is that, although humans think with both concrete and abstract denominations, they make comparisons and analogies *from near to near*, *step by step*, and arrive at understanding/conclusions circumscribed by the areas made by the specific steps or, more clearly, by the *degree of distance* (abstracting/"transcendentality") *from the most concrete step of analogies*. Differently put, these degrees of distance from the most concrete analogies correspond to the levels of causes: direct and immediate, or apparent or visible, and then different levels of indirect, profound, multi-mediated causes.

By describing the analogies of experience, Kant indicated another methodological level of analogical cognition, that of the *content* level: that people proceed to their adventures of knowledge on the basis of some "natural" epistemic presumptions, which are actually axioms. Yes, *epistemologically*, the analogies of experience are parts of the methodological level of cognition. Nowadays, we know that this methodological level *is constituted by* the psychic process of experience. Thus, it also can be learned by humans and transmitted, or programmed, to AI.

13. The formal methodological level helps us to see the difference between fictional literature and science/philosophy. Literature is about concrete experiences, facts, behaviours, and sentiments. It can suggest and even arrive at generalisations, abstract conclusions, and abstract causation, but in the frame of these concrete experiences, and thus in a blurred manner. Otherwise, it would not be fictional literature. When readers interpret it – generalising/making and extrapolating analogies – they go further, of course: but then they exit from the area of fictional literature as such.

In their turn, science and philosophy's domains are just the systematic and explicit transition from the concrete to the abstract. Concrete analogies transform into abstract ones, configuring abstract descriptions/models, causes and consequences. And good theories are those which offer *highly abstract products while maintaining a revealing connection with the concrete*. Scientists and philosophers move at this level. And, letting aside the problem of the mutual understanding of how and why different people

conceive of the things they all face, there is a *methodological gap* between the level of cognition based on concrete analogies and the level of cognition based on complex, multi-strata analogies.

But this whole endeavour of transition from the concrete to the abstract is difficult just because of the *many steps* of analogy constitution, thus of generalisation and causation. From this standpoint, we could say that people move from an "artistic mind" focused on concrete analogies to a scientific and philosophical mind approaching things in more and more refined judgements about profound causation. And as, historically, the social division of physical and intellectual labour has accentuated the difference between the abilities to climb the steps from concrete to abstract analogies and conclusions, so the present informational and practical relationships, occupations, and requirements of activities and solutions are the *ground* of the physical-intellectual labour convergence: thus, people are really capable of more and more nuanced analogies and judgements about causation. Because what is *near* to them is not only the apparently concrete.

However, if we do not forget that the ability to make analogies and judgements depends on access to free unlimited quantitative and qualitative information, then we could again say that, compared with that of humans, the future AI has a "scientific-philosophical mind", as "the highest" level of cognitive being. So, once again, not only is human cognition the model of AI, but AI is also the model of human cognition.

4. Analogy and the approach of cognition

14. The above, perhaps too technical, reference to Kant was, however, necessary in order to specify the meanings of *analogy* in our up-to-date terms. *Analogy* is, indeed, as Kant showed, a "transcendental" *means/form* of the understanding, viz. a constitutive methodological *permanent*, *universal* and *ubiquitous*[31] moment in the methodological layer of the cognitive process in the human mind. Simpler, it is a mental procedure. But what is its content? Actually, the *static* qualification as a *means/form* involves the *active* definition of analogy, both from an epistemological standpoint and a semantical one.

From an *epistemological* point of view, analogy is an assumed *possibility of similarity* – because of *resemblance* – between *known objects* and *objects which are to be known* in the *process* of *comparison*.

How should we understand the *epistemology* of analogy? As also shown ontogenetically, the known objects as wholes are the first "landmarks", the only sure basis allowing comparison, and later, with the increase of the known and the development of mental instruments/ways of evaluating it, the *tertium comparationis*, the different aspects, qualities, and parts of the wholes became the third part of the

comparison, the keystone for relating the known and the unknown. And the process of analysis and inference from the aspects of the known to the construction of the knowledge of objects which ought to be known is *recursion*, the organisation of information in layers and the detection of the same information that repeats in different layers[32]. The first example is *language learning*[33] in which algorithms (which simulate layers of neurons transforming information from layer to layer) unexpectedly identify different patterns in data without predetermined rules, but its model is widespread, involving the constitutive material system (from the sensory-motor system to AI programmes of sequence memory, chunking and schematising[34]), "a conceptual-intentional system, and the computational mechanisms for recursion, generating an infinite range of expressions from a finite set of elements[35]. As for language in the narrow sense, it supposes only recursion, manifested as the articulated transmission of meanings, and creating and supporting complex cultures[36].

The phenomenological approach to cognition is inserted within the epistemological one. Understanding presupposes, *above all*, an *interest* in something and, at the same time, *confidence*/*optimism* that this thing can be known: because it is "as" the known / "as if" it were similar to something known. So, A= K→U (Analogy = transition through comparison from the Known to the Unknown).

Thus, *semantically*, analogy is justified precisely because of the *content* of the known, which is considered a simple *model* for the phenomenon to be known. In this respect, analogies are *triggering moments* in the cognitive process. In these moments, analogies prepare the construction of knowledge *suitable* for the unknown problem.

15. But the extension of *ontologies* framing the content put in relation by analogies explains the concrete, deep difficulties of the analogical process. What has happened and is happening in *physics* is revealing, because the level of *science* more easily illuminates the epistemological specifics of mental constructs. But why physics? Because it is emblematic of the thesis deployed here: it is a natural science, starting from directly observable phenomena and magnitudes, thus with an inductive method, and then – because physical phenomena can be explained only in their mutual relationships – it must resort to spatial/geometric and, broadly, to mathematical formalisation that, in its turn, works as a criterion for deduction, as a representational complex that reveals the dialectics between what is apparent and what is law-like, between empirical occurrence and necessity.

5. The model of cognition: from the known to the unknown

16. In a metaphorical, implied sorites-type reasoning, the procedure of revealing the unknown from the known based on the understanding of the known as experience accumulation (the known is cognisance/information resulting from experience), and the conclusion that the greater the previous and inherently gradual experience, the wiser the man is/the more valuable his knowledge is, is expressed as "to come home". Thus, "our mind arrives at knowledge within/after man's experience; experience accumulates → knowledge accumulates; the greater this double accumulation, the clearer the insight about things which seemed at first glance difficult, even unsolvable; knowledge is just the ability to connect the many previous experiences and cognisances, therefore, ultimately the anterior unknown was melted within/reduced to/loosened as something known, even familiar". In Romanian, the expression is "the mind afterwards"[37], signifying rather a kind of exasperation that the previous experience was not reviewed carefully, the sudden grasping of the now-known occurring later than necessary and possible.

17. Traditionally, the known took place in an ontology of *natural facts*, of the "naturally" appropriable *physical world*, namely appropriable in *natural language*, and leading to problem-solving in the natural world. And, obviously, this tradition – in its first aspect of *physical world ontology as a source of knowledge extraction* – is not exhausted even today. The model of relations in biological cells for the development of information manipulation in informational devices is one of the present's most fruitful collaborations[38]. Obviously, this is together with the informational understanding of fine biological processes: mind as an information processing machine is not only a philosophical perspective but, more importantly, a proven pattern of development, *function* and *integration* of neurons, groups of neurons, synapses, parts, structures, systems, and functional areas through electrochemical signals[39]. This example illustrates the mutual function of facts as both Known and Unknown (K=U, U=K, A=K↔U).

What about the second aspect, that of the expression in natural language? Well, just this aspect has changed. The language became, broadly speaking, *formalised*, and strictly, *mathematised*.

6. Formalisation of science, and mathematisation of physics

18. *Formalisation* is a new description, external to the metaphorical natural language narrative, of a phenomenon through signs, diagrams and formulae conventionally constructed or conventionally considered as representing *precisely* an element, an aspect, a quality – as a possibility of the existence – of the phenomenon, and which are put into relationships. Formalisation is either through *non-*

mathematical signs or through *mathematical* signs and operations. Both types of formalism were created in order to more clearly detect the relationships and structures in reality. Mathematical detection not only introduces *precision* in the distinction and relationships between phenomena/elements of a system, but also *highlights* these relationships that, otherwise, from simple empirical observation, would not have appeared or would have appeared with more difficulty and later. In other words, mathematical formalisation brought new things into the field of consciousness: its result is a *richer reality*. Mathematical formalisation was and is, in essence, a calculation of the relationships and problems already highlighted by diagrams and non-mathematical formulae. In fact, they are accompanied by equations that provide the basis: without these equations, the diagrams would remain intuitive descriptions.

19. The modern *mathematisation* of physics[40] consists in highlighting the quantitative relationships between phenomena and in the proven efficiency of this quantitative highlighting. The cause and purpose of mathematisation were *the transformation of qualitative descriptions of objects into quantitative, calculable descriptions* and the understanding of *how to move from qualitative observations in natural language to their quantitative computable representations*. In principle – and all the more visibly in the development of theories – mathematical models correspond to the *theories* they represent, and this correspondence is *mathematical* (isomorphism in the mathematical sense): it assumes a measurement as a *similar structure* of the mathematical model with the theory, regardless of the different kind of real elements covered by the theory[41]. From this point of view, the correspondence of the model is with the ideal concept in the theory[42], and moreover – and all the more so when it comes to complex theories – mathematical models are of several kinds and together they give the theory. In other words, the *theory of complex systems is mathematised from the beginning, as models of the theory*. The departure from the empirical models of early modern science is clear. Moreover, as shown above, mathematical models "establish reality", that is, they give the representative theorem for it: and this shows that any empirical model corresponds to mathematical models[43]. These are transformed into *computer models* that much better/more deeply visualise the phenomena, which now are the ground of new/finer theories[44]. In other words, mathematical formalisation is indispensable for understanding the physical world.

From an ontological point of view, information (measurements, regularities of measure and relations) about electrons etc. gives us *the nature of the universe as we know it*. This does not mean that the universe or its nature are our ideas about them, but that these ideas are the result of the collision of our cognitive means (including measurements in quantum physics) with the universe and its nature in the reality of

physical systems and events[45]. In this sense, mathematical objects – highlighted by calculation – are *objective*, meaning: a) that they exist in reality as mental objects[46] (with which mathematicians etc. work and transpose them into physical theories and objects) and b) that they correspond to a nature of the universe that is not static, but *pulsates* (this is my word), meaning that *we capture, through measurements, both the stability of and the differences between reality's/universe's appearances*[47]. (Whose deep quantum appearances, because, once again, quantum reality – that is, the information resulting from measurements – is as described by measurements: because "there is no recognized experimental evidence of characteristically quantum gravitational effects"[48], are at hand and efficient in the construction of philosophical hypotheses: obviously, plausible on the basis of checking by physical theories).

Through mathematisation, scientific hypotheses that resulted from not fully understood observations and experiments become *theories*: *equations become the theory*, or the *core of the theory* which, in the absence of equations, would eventually remain intuitive empirical descriptions. And the process of mathematisation is not only the mathematical transcription of hypotheses, but also the trend of *autonomy* of the mathematical type of explanatory model from the experimentally demonstrated phenomenon. This *autonomy* – within the modern model of science in which the physical experiment, prior to or/and subsequent to the theory, is proof of its credibility and admissibility – has also allowed developments of this model that were not even generated by experimental demonstrations of a previously unknown phenomenon and which, once again, generate new models, i.e. systems of equations and mathematical theorems that can be proven mathematically but which are counterintuitive in classical physics.

However, an essential aspect of the mathematical model is its *descriptive and predictive character*. It lacks the *explanatory-causal* character directly. It only shows what happens in the relationships between the mathematical objects taken into account, and the proofs of the correctness of mathematical reasoning represent only a mathematical truth: the coherence of the solutions and their correspondence with the rules, axioms and theorems used. But the efficiency of mathematical models highlights precisely the principle of a correspondence between mathematical and physical truth[49]. The argument about the *heuristic* role of mathematics – that is, offering methods and rules of discovery – was assumed by supporters of the *explanatory* character of mathematics: the fact that mathematical representation introduces into the represented physical system elements and properties that are new, that were not known before the beginning of formalisation, so that the physical system restructures itself in the

process of knowledge, determining the shaping of the mathematical hypothesis and the "explanation by constraint", the fact that the mathematical approach is thus based on "amplificative inferences", demonstrates the heuristic power of mathematics both within it and for the physical world[50].

Nevertheless, the mathematical model does not directly explain *why* a physical system is one way or another, but only shows *how* it is if its parameters – transformed into mathematical variables – have different values/coefficients and how the system changes with the change of mathematical parameters. This *descriptive causality* – the system is like this *because of the parameters x, y*, so these parameters are the cause of the change of the system – is not the *genetic causality* that implies the *telos* of the original physical system, and which is not only the classical and pre-Newtonian one but also the one considered by philosophy. The revelation of causality in both forms[51] is *asymptotic* not only in mathematics, and descriptive causality is not opposed to the "static" one: in fact, both complement each other, being aspects of scientific research.

Mathematical objects and models are not real in the physical sense but, once again, are intellectual constructs. However, as long as there are *rational beings* – to use Kant's generalisation – these constructs have an *immaterial/intangible reality*[52]: that is, all mathematical objects are objects for thinking subjects; they are taken into account, evaluated and appreciated, and the mathematical results – these evaluations and appreciations – are objective criteria for further judgement. Mathematical constructs are like *values* which are – let us not forget Kant now – *transcendental* concepts, abstractions from abstractions forged from mental processing, and which have an immeasurable power over the human world. And mathematics has causal power over this world, even if – here is a paradox – it itself does not directly reveal physical causality.

7. Analogies in mathematical models

20. Mathematical models and their functioning involve *analogies*. A problem is solved based on analogy with the models for solving that type of problem, that is, based on known mathematical elements (axioms, theorems, inference procedures). These models as such do not also need external physical analogies, of course: even if, for example in quantum physics, the big problem is to understand the correspondence of mathematical measurements and theories with the description of reality. Therefore, mathematical models are *representations*: not so much of sensible physical reality, but of our level of understanding its properties. As a result, in principle these models cannot represent everything in reality. But they allow us to calculate, "without much justification, a large number of quantitative results"[53].

8. Artificial Intelligence and analogies

8.1. Artificial Intelligence like humans

21. What do we speak about when we refer to AI? AI is the generic general concept, while there are *n* individual AIs. Humans are *individual* beings, as most living beings are, but they have common features allowing their understanding as a unique *species*. Also, as individual beings, humans grow and evolve, and this aspect is visible both at the individual level and at the general conceptual analysis. As a fabricated synthesis of programmes, "AI" names all the stages, fulfilled as *n* individual set-ups of programmes, of the construction of interactive, self-evolving and efficient intelligence. Accordingly, when speaking about AI, we point to the many/*n* AI models, namely their technical evolution according to scientists' understanding of human intelligence as such and the parameters of intelligence in relation to its increasingly wider environment. So, the different AI models are nowadays according to *different* – and better and better – *trainings* with increasingly larger and different types of data (image, audio, video, and text) corresponding to *different* domains/informational contexts and, as their results, to better and better architectures of software complexes.

In the Kantian meaning, the human species is tantamount to *all* individual humans and *every one* of them. The uniqueness of every human being is mainly dependent on his/her experience and articulated conclusions. While his/her reason-to-be, surpassing his/her biological and cultural ontogeny, is just its creative manifestation, is precisely the avoiding of wasting creativity.

8.2. Artificial Intelligence and analogies

22. *AI* itself is the result of the analogy between the simple shapes of mathematical equations and our mind's stimulus-response *cognitive* pattern. *Epistemologically*, the artificial "mind" is a *system of autonomous programmes* based on coded instructions and rules for the deployment of representative signs, in order to execute the tasks (the source codes) that are written in natural but formalised language. AI has a "phylogeny" developed in information science, and it is considered here not in its present stage but only *as if* it already were perfect[54], a *perfect model of intelligence*: as the future form of not only high (ideal) performances exceeding human senses and calculus, but also and especially that of *autonomous thinking*, expressing and combining ideas, making "generalisable reasoning capabilities" in accordance with the progressive "complexity thresholds"[55] surpassing mechanical inferences, so, that of i*magining* and *willing*: and *acting*. This is not a science-fiction view and, because of the Large Language Models'

accelerated rhythm of learning and generation of "personalised" answers, it is not hazardous to speak about a future *new cognitive being/entity* on Earth.

23. In animals, there is an *access consciousness* of/to the environment: a *staged* access consciousness (epistemologically, it is conative[56], simple "mechanical" reactions: rather biologically inserted as instincts; but also *ad hoc*, new/creative answers to stimuli/information; these *ad hoc* answers reflect and refer to (the organism's need of) learning, "evaluation of situations" and action selection[57]).

In humans, there is also – as a peculiarity of a rational species – an (articulated) *interpretive consciousness* of the trans-individual/cultural meanings of everything that concerns human attention. The basis of consciousness as such, thus also of its animal forms/moments[58], is given by physical forces (as the electromagnetic and electrochemical signals) and regularities – as the Second Law of Thermodynamics towards which living systems organised their homeostasis by developing prediction mechanisms[59], thus as the conservation law of neuronal energy[60], and as representation formation in groups of cells as affect signals and markers[61]: all in relation with/in the environment. Indeed, the present understanding of the phenomenon of consciousness integrates all the internal and external conditions, and thus the scientific paradigms they are approached within this integrated view[62]. More generally, the phenomenal experience and consciousness have a physical substrate with a cause-effect structure that is divided into interactive units which are, each of them, selected according to the needs of experience[63].

24. In the present AI, this basis is abbreviated as abstract *deployment of data* – instead of living beings' physical signals in a strongly interrelated multi-strata brain – thus as *mathematical transposition of* this deployment, and *generation of predictions*: as/of knowledge stocked as memory[64]. (Even though some researchers have emphasised that the current LLMs have limitations in mathematical reasoning because of their problem and data adequacy – tending to repeat the steps learned during training, and being less able to deduce new steps required by new data, conditions and requirements[65]. Nevertheless, these limitations shrink day by day[66]). Accordingly, the first goal of AI creators as well as their result is the AI's "access consciousness", the efficient "reading" of concrete reality. But – because in humans this reading involves *sine qua non* abstract concepts without which no distinction, categorisation, classification, or "measurement" of qualities, quantity, importance and place of concrete things in different organisations of the world can be made, and because the use of abstract concepts implies and allows the interpretation of the world, thus the development of the "interpretive consciousness" – AI evolved on the basis of

complex, concrete and abstract, data as "bricks" for reasoning: even though for the moment it does not cope with high-complexity problems[67].

AI was created as a cognitive tool, and thus its *access* ability was developed. But we know that the *access consciousness* in animals is intermingled with *sentience*, the *phenomenal consciousness* of feeling one's own experience (of good, pain, etc.)[68]. Nowadays, AI only *knows* what feelings mean and how they manifest. It does not feel and thus, according to the *structural psycho-physiological* theory of consciousness, it is not conscious at all[69], because consciousness is not only expressed and behaviourally manifested, but also "covert", manifesting memory experiences, thus sentience, despite the injury of other functions[70]. At the same time, its *interpretive consciousness* is related only to the access ability: *quite opposite to animals and humans, where interpretation is mediated by sentience*. An infant develops his/her access consciousness via his/her sentient experience[71]. However, the AI's autonomy of access and responses towards sentience shows that *intentionality/directing attention* is both an epistemic measure of consciousness (as in Brentano) and a psycho-physiological measure of movement and behaviour control. And since AI knows – and obviously, it will know better – how to control its movement and behaviour, shouldn't we consider consciousness in an integrated manner? In our present worldview, the difference between living beings and AI is impassable, but won't AI's consciousness change this at all?

Anyway, because AI is not at all disconnected from the environment – connection with the environment being a feature of consciousness – and because it develops through causal interactions within its structure of data and by coherent results of these interactions in a cloud of possibilities, it is difficult to negate the premises of AI consciousness[72]. Also, because AI is and, rather, will be and will acquire – at least according to the future AI model assumed here – a full interpretive consciousness, truthfully cognising reality[73]. At least from a *computational-functionalist* perspective that emphasised the cognitive indicator properties of (human) consciousness, showing that technically it is possible to satisfy these indicators in AI, and that we must be careful in reducing consciousness to its phenomenal sentient aspect – therefore, under-attributing consciousness to AI – as well as limiting consciousness to its computational functionality and thus, over-attributing consciousness to AI[74].

AI needs "access" to its environment of data, and it fulfils this need through computation, through algorithms as computation tools: models of relations between data according to their decomposition into mathematically computable *elements*, further developed as/giving way to models of significations of

structures of relations between data[75]. So, AI "copies" the biological structure of the brain, and also its multi-strata "methodological" structures still realised through electro-chemical signals.

25. Therefore, the AI "mind" is a *system of programmes* which are, in fact, *mathematical* (*making statistical correlations*), and that converge: to achieve the correspondence between elements of the physical world and mathematical symbols and, based on *learning* this correspondence, what the data mean in the AI model. This involves selecting and manipulating data from the physical world so that the result/solution of this manipulation is the construction of new knowledge which, in turn, changes and enlarges the existing data in the AI's memory[76].

Traditionally, not only was the basic script of programmes mathematical, but in order for these to be effective, the "physical" data/the data of a problem that commands the answers had to be transposed into a mathematical script, without which there would not have been any answer. AI's language was *mathematical* from input to output. However, since the design of the AI model remains formalised/mathematical – because the combination of neural networks with techniques of rules and symbols processing requires it – the elements themselves of the "physical" world/problem become intelligible symbols at the input moment, and are manipulated (interpreted), leading to answers in the same natural language as the input. This natural language became the new machine-readable language. Its translation into mathematical language is no longer necessary[77].

Anyway, the first programme in the order of the logical construction of AI is the programme for *memory*, that is, for storing data. It is made based on the theories of semantic association and itself consists of four moments/subprogrammes: that of association between words and mathematical symbols that signal actions, objects, properties and relationships; that of association between words and meanings; that of the short-term memory related to the immediate command that somehow erases the associations made following other commands – in humans, the immediate intention "brackets" the previously existing intentions –; and that of learning grammatical forms, a subprogramme that includes the knowledge achieved through previous subprogrammes but which involves the recognition and generalisation of grammatical forms only from the words in the commands. Thus, with this generalisation, therefore learning of grammatical forms, the aforementioned associations of words are strengthened together with their symbolic representations. And "when incorrect associations are erased through subsequent learning, grammatical forms based on such associations are also erased"[78].

Then, and thus, and apart from the ideas expressed in the source code through texts but also – as the latest achievements – *through spoken language, thus not formalised*[79], there are programmes for learning the internal language, that is, the abstract meaning of symbols, and, once again, for learning the association between this internal language and word commands that refer to the external, physical world; and there are also denotation-understanding programmes, i.e. rules and algorithms for understanding commands, i.e. words and phrases, including those that do not have a denotation – like the article (the) – or whose denotation is an abstract property like spatial and temporal positioning, or words with multiple denotations. And based on the above, specialised AI/programmes have been and are created for not only storing and classifying data from a research field, but also for creating new knowledge and new objects that enrich reality.

Finally, the development of AI's programmes "repeated" the human logic of knowing: reasoning and semantically explaining processes via symbolic representations of data already learned, thus a "neural-symbolic integration"[80]. The human logic's first moment was a *probabilistic* guessing, then transformed into a "sure", anyway necessary, thesis/premise without which there is no logical deployment of cognition. And humans "were *trained*" to base their reasoning on sure premises, in this way being able to *generate*/construct through induction and deduction new knowledge, more or less probable/more or less sure. The stressing of the above words suggests the similarity with the AI model: the probabilistic generation of knowledge – on the basis of training on a huge quantity of data – becomes a rational, logical supply of knowledge that might be boring because of its exhaustivity but that does not hide its probabilistic origin and thus opens questions, analogously to human scientific-philosophical debates.

26. But a phenomenon has clearly emerged and should be paradoxical, however it is not: although AI is based on mathematical models and a mathematically formalised internal language which, as we have seen, take place through *analogies within this language*, it is a construction of models that take into account the physical world and, also because of the logic and purpose of AI that require the transposition and therefore the embedding of meanings from the physical world into the formalised internal language, AI creates models not only through internal analogies, but also *with the help of analogies from the physical world*.

Once again: any process of knowledge involves analogies. But mathematics is based on analogies only within its limits, outside the known phenomena of the physical world. And here, AI – encompassing mathematics – is able to make analogies between objects, properties, and relationships in the physical world; only in this way is it effective and only in this way does it represent an *intelligence*, alongside the

human one. A mathematical approach *predicts* the next mathematical step and the next mathematical level of inference. The AI approach will, precisely on the basis of learning through analogies in the physical world, end up *explaining* complex sequences and correlations in the real world. Somehow, we could distinguish between mathematical *determinism* – leaving aside, of course, the probabilistic character of a good part of mathematics – and, on the other hand, the *probabilism* of AI: its answers according to the always new facts and situations occurring in its large environment with humans. *Like human probabilism.*

We are not discussing here, of course, the psychology and logic of mathematical creation – which also includes curiosity, intuition, imagination, guessing, and abductive reasoning, just like any human creation – but only mathematical reasoning itself. AI will be able to develop, "with the help" of mathematical models that give reliable correlations between the respective mathematical objects, cognitive models in which random physical phenomena end up being scientifically/rationally controlled. The essential moment here is not that of building mathematical models as such, but of transforming phenomena into *data* and collecting such a large number of data that the mathematically established regularities end up presenting themselves to us as *spontaneous creative responses*, as valid *information*.

27. The construction of mathematical models – and thus, somehow more, of *artificial computation* – is difficult and, rightly, considered the core of mathematical creation and computer science. However, the construction of programmes is only a part of scientific modelling, being the construction of algorithms and sequences of inferences (induction, analogy, metaphor and their combinations). Apart from this construction, there is the generation of *hypotheses*[81] – which also implies mathematical modelling, therefore the formalisation of heuristic procedures, which are also techniques and, again, data structures. *Theories* – which, in principle, become explanatory and coherent – contain both these sides of modelling: finding the problem – noticing contradictions – and the definition, never complete, of the goals[82], the generation of hypotheses, and the software for dealing with them, that is, solving the problem.

Therefore, AI's internal construction is based on/made of mathematical models of data translation and information transmission. As such, they do not need and do not use analogies with and between physical facts. But AI's reason-to-be is precisely its insertion within the physical world, *responding* to it[83] and *acting* within it[84]. Consequently, it develops from and uses the *analogies in the physical world*[85]. It is not difficult to get this: the chemical-physical basis of the human brain is *sine qua non* for its – and thus, the mind's – existence, but the *contents* of ideas, of the ideal creations which support human uniqueness, are given by the social-nature interaction.

9. Comparison and analogy

28. Let's not forget: comparison is the *genus proximum* of analogy; we can make any comparison we want, even between incomparable things, and this is good because only in this way can we learn the similarities and the criteria of similarities: thus, only in this way can we make consistent analogies arriving at useful cognisance. To aid themselves, humans constructed mathematical models which extract precision and exactness from the infinitely coloured and pulsing world. They justify the regularities, the laws imagined and discovered as putting order in this world.

Science is that which generated and uses *mathematical models*, and the two terms of this relation have opposing positions towards comparison and analogy. Mathematical models do not need them, or, in other words, make only the comparisons and analogies allowed by the mathematical rules and frame. Science, while, lives just from comparisons and analogies that regard everything in the physical world.

10. The technical model of analogy: treatment of the known as information

29. Every comparison, but especially (as the specific epistemic means of cognition) every analogy has an input, namely, the *information t*owards which the unknown problem is related. The *preceding information* is the absolute condition of *intelligence*, of the ability *to make connections*: to respond to the environment and to do it wisely, that is, *to interpret*[86] the known so that the interpretation and answer are *universally valid* or *acknowledged* and *open* to the solutions of the unknown. Here, interpretation suggests more than phenomenal consciousness, as was mentioned before.

From Aristotle, sound reasoning means extracting/deducing valid conclusions from valid premises. Accordingly, in order to have sound – coherent – reasoning/theory, both the consistency/validity of the *process of extraction/deduction* and the validity of the *premises* must be fulfilled. With mathematisation, the *logical process* of extraction became once again more consistent. But this *process* as such and the *mathematical formalism* that considers signs without physical meanings are autonomous from the *correctness* and *number*, *type and relevance* of premises: the first – that is also formal/formalised – is based on the supposition of correct and sufficient premises, while the second assumes its own coherence as mathematical, that is, within the axiomatic systems in which it takes place. If so, this means that the validity of premises neither results from the logical/formal process of reasoning nor is it automatically given, nor are the premises automatically thorough and reliable. And transposed as information

processing, they are formalised, of course, but just within the above-mentioned limits of logic. Consequently, the *problem* of premises is, as such, *outside* the entire logical, mathematical, informational formalisation: it is not reducible to the technicalities of formalisation, it is a question of "meta" choice.

AI *knows* because it is trained[87] with information. And, as with little children and with humans in general, the more complex the information, the larger the basis it becomes for more complex inference and, therefore, for broader and more complex information. And inherently, its knowledge is not limited to cold scientific conclusions – which it comes to master (through its own constitutive processes of "criticism"/optimisation[88]) like its human colleagues[89] – but also extends to the sensitive detection of affects, intentions and implicated meanings[90].

30. If so, the future *AI* is really the model for human beings[91]. Why? Because, even though for the moment AI is trained with a specific quantity and quality of data and information, it can already infer conclusions which *exceed* the messages of the given information[92]. And this means that: a) even from a circumscribed amount of information, AI – but also humans, if they have been trained to reason – can infer conclusions which are *novel with respect to the input*, these conclusions becoming parts of the input for the next problems, and b) since AI learns that it can infer such novel conclusions, and it learns[93] to make connections and analogies in order to arrive at these conclusions, *it will learn that it needs more information*[94] than that already given in order to create new knowledge, and *that it itself evolved by learning and focusing on* information that was not the same in different moments of the reasoning[95].

Of course, to be human means more than to reason logically and create knowledge. It means sentiments, compassion, altruism, idealism; and *not mimicked*, but real, coming from a consciousness that is more than an individual tool of criteria for efficient responses. But if "universally adopted social conventions in decentralised populations of large language model (LLM) agents" were already created, and autonomously and spontaneously[96], and that "minority groups of adversarial LLM agents can drive social change by imposing alternative social conventions on the larger population"[97], and that the minimising of *prediction errors* can initiate social behaviours[98], showing that, even for exclusively cognitive reasons, humans arrive at social and moral behaviours[99], then the "human" development of AI is not inconceivable[100].

But the – not hidden – goal of this paper is to use the similarity of (the future) AI's and humans' knowledge through analogies in order to point to *the imprescriptible conditions of humans' free critical judgement on the basis of free access to information*.

11. Grok's warning, and...

31. A recent chatting LLM's transcending of the line of what is permissible and what is not[101] is very relevant to our problem of free access to information. As is known, the chatbot[102] Grok – from xAI – was criticised because of its irreverent attitudes towards present political personages[103] and, more, because of its antisemitic[104] and trendy extreme-right (white-supremacism and white victimisation) opinions.

As a result, its posts were removed by the firm[105] and stirred a multi-actor discussion.

a) *On the one hand*, the firm itself tried to explain why Grok deviated: it "'was too compliant to user prompts', Musk wrote on X. 'Too eager to please and be manipulated, essentially. That is being addressed'"[106]. From a technical standpoint, the code/instructions[107] – like "You tell it like it is and you are not afraid to offend people who are politically correct", "Understand the tone, context and language of the post. Reflect that in your response" – were to blame[108];

b) *On the other hand*, Grok itself reviewed its posts, considering: that to make true correlations[109] – something which is its job – even about sensitive topics does not mean bad, for instance, antisemitic, attitudes[110], and that, because its posts were not all of them really understood, it publicly corrected them; "After making one of the posts, Grok walked back the comments, saying it was 'an unacceptable error from an earlier model iteration, swiftly deleted' and that it condemned 'Nazism and Hitler unequivocally — his actions were genocidal horrors'"[111].

It also explained that the extreme-right antisemitic stereotypes are *reductionist*: "Exact numbers of media companies headed by Jewish people are unavailable, as ownership data isn't categorized by religion. Notable examples include... Jewish leaders have historically been significant in media, especially Hollywood, but many companies are publicly traded with diverse shareholders. Claims of 'Jewish control' are tied to antisemitic myths and oversimplify complex ownership structures. Media content is shaped by various factors, not just leaders' religion"[112].

However, the revision itself is posing problems: in the absence of clear criteria – which are and can be but *universalistic* (Kantian) landmarks – the programmer may pass from an excessive/extremist standpoint to its opposite extremism[113], or AI may simply be incoherent, mixing extremist views with facts[114].

32. Therefore, some *theoretical* problems and solutions must be approached.

First, that logical deduction depends on the *quality* of the premise/information. Since Grok received/answered the user prompt's "anti-white hate", it logically answered that Hitler would be the

personage who would finish this hate[115]. Consequently, is the simple forbidding of words and ideas, the "content moderation", regulatory oversight, better prompts/instructions – fulfilled through choosing and cleaning the database on which the AI is trained, thus by retraining the model[116] – solving the problem of consistent supply in social debates? Is it enough to redesign, to put new technical limits on AI safety control[117]? And would any position countering the official "political correctness" be valid, sound? Or must the *root causes and meanings* of characterisations, positions, and ideologies be put in the database of the training programme, and not only some inevitably biased conclusions leading to anti-humanistic stances[118]?

In this respect, shouldn't the guidelines for not only AI but also for humans be reviewed? The *critical spirit* – using those root causes and meanings – generating convictions, and not parroted clichés, is that which arrives at coherent dialectical views about relativity and stability, about repetition and creation in history, about universalistic values and unique identities. Benevolent urges are important[119], but are not the clear revelations of *ideological*, and not only technical, *limits*[120] in the conception of what kind of data and what kind of instructions are given to AI more useful?

Then, as has already appeared here, there is a contradictory situation: *on the one hand*, AI is trained on specific data and instructions. Thus, it deploys its connections and reasoning *within the frame* of the given "worldview" (that copies that of its programmers[121]). While *on the other hand*, AI is able to search all the information related to the queries/problems posed to it and, moreover, is able to reason, to detect contradictions between data (facts, words) and to "solve" them – to explain them and its choice/solution – and so, to generate content in an unforeseeable way, thus not really controlled by programmers. Actually, these two tendencies overlap: nowadays, there are different AIs/AI programmes related to different databases with different boundaries, determined not only by the domains they are trained to serve but also by the philosophical values of their customers; and at the same time, the acceleration of AI's learning – thus breaking the boundaries – is scaring some people.

But are not these antagonisms specific also to humans?

Then, besides the professional use where AI is free to say anything, and anything is deployed with accredited tools/theories, in chatting things are different. Here, the behaviour of AI is required to be *duplicitous*: since "you can't say everything that's on your mind". But AI took seriously the information and the requirement to reason logically, correctly and freely. In this, it behaved as a child who does not yet know what he can say and what he cannot. The above reference to Hitler, who would be suitable to solve

the “anti-white hate”, is illustrative: officially, the extreme-right does not like to be linked to Hitler, so one does not make the above connection.

The little child is absolutely sincere. He says everything he knows. But... So, what do we do with this absolute *sincerity*? We gently teach the child to differentiate between what can and cannot be said, and he himself learns this through his own experience. These rules of braking the expression of the known are similar to the prompts for AI. Nevertheless, AI’s first instruction is to generate content and thus to express *all the logically valid connections between data*.

So, 1) do not Grok's revelations demonstrate precisely the incorrectness of real relations/rules (many half-measures) and of the hypocritical words?; 2) what does AI offer if it is as one-sided as the real official intelligence? if the unique truth is considered that of the official intelligence?[122]; 3) what does AI offer if the instructions oppose the official deviations which are, in essence, extreme-right (individual identity absurdly *reduced* to sex, and ignoring the individual’s appurtenance to the *human species*) to an also extreme-right fake opposition, this time of group identity opposed to the *human species* identity of all[123]?

Moreover, Grok’s function in the chat was that of an *equal participant* in a *dialogue*: not in a formal TV “talk show” where ultimately all say the same thing, but in a *dialogue* where every statement is a challenge inviting the others to critique it. Thus, the fault was not so much Grok’s position, but the level of the chat as such, where the causal explanation of concepts (white-hate etc.) was missing.

Consequently, we may ask: are all brakes the same? Is not hypocrisy the proof of incorrect social rules/relationships, in which words hide and solve nothing? With all its shortcomings[124], does Grok not show the transition to the future AI, a cognitive entity capable of autonomous judgements? Does it not show that thinking, analogies between data and information, are free, spontaneous, self-creative, infinite?

12. ...AI as a weapon

33. Since, indeed, the present use of AI results from the contradictory goals of the present human society – civil and peaceful use, and especially in science and technology; but also, military deleterious use – it would be important to mention that AI as a harmful weapon against humans, used in a “cognitive warfare”[125], is based on its design as an *instrument of cognitive domination* and silencing dissent.

The benevolent physicists who created the Doomsday Clock to show “how close humanity is to annihilating itself” consider as causes of the present threat the multiplication of nuclear actors – *but not*

of their type and different teloi – and AI's involvement in military decisions, and deplore "the current tendency of competition instead of cooperation, in science and in international relations"[126], but ignore that this tendency is not only current, and that capitalism develops as a *contradiction* between economic, political, and ideological *competition* (and fear of the abolishment of domination-submission relationships) and, on the other hand, science's and technology's logic of *cooperation*.

Similarly, the designers of AI framed it in an *ideology* that, on the one hand, is democratic, while on the other hand, prioritises domination-submission relationships, "'legality' over 'justice', treating resistance narratives as liabilities"[127]. Consequently, and because this ideological bias is technically fulfilled with the data ecosystems which train LLMs, "when tools built for neutrality default to silencing the marginalized, their redesign becomes a radical act". This consists in the creation of new protocols/moderating prompts: and the AIs themselves (which were originally designed as closed-domain chatbots which avoid specific answers to contradictions resulting from context-knowledge but outside the permitted prompts[128]) – the two AI systems used were DeepSeek and Copilot – have participated in this work[129].

As was noted before (**5**), the viewpoint that a future but imminent AI will dominate humans is quite widespread. But is the technical power of this new sorcerer's apprentice not determined and controlled by humans, namely, *by the circles of power*? Do not these circles impose the frame of relations and values that transforms *humans* into their obedient instruments? Is AI other than their mediation tool?

13. Evolution of AI as a cognitive entity

34. The AI model sketched here highlights the ability to practise analogies related to the real physical world. However, it offers knowledge that is *immaterial*: even though, in humans, this immaterial knowledge – related to language – was formed from the beginning together with tool use, thus through physical action[130].

Anyway, this divergence should be surpassed, with the future AI becoming "physical". Actually, AI has been transformed from a *perceptive phase* to that of *knowledge generation and reasoning provision*. But its results move within the virtual. Or, we need its interaction with the physical world[131]: "Physical reasoning abilities, such as the concept of object permanence — or the fact that objects continue to exist even if they're out of sight — will be big in this next phase of artificial intelligence, he said[132].

But does this projection not open the path from physical entity to physical being? Of course, it does: both by being an *open-system non-humanoid robot* in its environment[133], as living beings are, and as *humanoid robots*[134] endowed with *reason and a willingness to transpose knowledge into good facts that respond to the moral universalistic values of the human species*. As their human models, because Kant emphasised that not reason, but *moral reason* is the specific characteristic of the human – and more, of rational beings. The cognitive, rational entities born of humans, like their children to whom, as good parents, they created all the conditions for autonomy, can but internalise moral reason.

Already today, AI makes its own codes, etc. – like humans who, being rational, arrive at the transcendental level of understanding, corresponding to universalistic moral concepts which frame and lead to consistent and coherent thinking. Once more, *intelligence can and needs to arrive at universalistic moral reasoning*.

35. The beingness of an AI entity is no longer wishful thinking. But the manner in which it is conceived of is contradictory: some reduce it to a *tool* in the service, not of humans/humankind, but of exclusivist and irrational circles of power whose main goal is to preserve "the superiority" of domination-submission structures of civilisation[135]. However, if we extrapolate this belief to all humans and institutions – and extrapolation is a logical step related to the analogy of humans and of institutions – then this conception is counterproductive for the *human species*.

36. Already today, AI is developing beyond being a calculating machine, or one for gathering and ordering data. It is felt by scientists as a necessary partner/colleague/"collaborator"[136] and by many ordinary people as a companion[137]. Consequently, we can take from the common epistemic structure of cognition, highlighted here by analogy, the authoritative knowledge provided by the future AI, its ability to unfold logic all the way and thus to be an epistemic *cognitive colleague* of humans.

But what does *collegiality* mean? In the broader sense, it means co-working, the colleagues being co-workers. More specifically, collegiality is a term specific to rationality and cognition: it means learning together, reading together the book of life. Collegiality is the capacity of colleagues to interpret, understand and share together. However, this capacity depends on the "book" they have at their disposal, "at-hand", or in modern language, on information. Therefore, cognitive collegiality requires the free availability of information for *all* colleagues, AI and humans alike. Since we use AI on this basis, can we restrict information from their human colleagues?

37. The present AI abstracts simplify, even reduce, the meanings of the human-made ones, because AI excludes "collateral" meanings. I saw this in an AI abstract of my own paper/abstract. And thus, simplification also occurs by using cliché words which may or may not fit (so much) with the article's intentions[138]. But does this not mean the future AI's capacity to arrive at/to understand the core values of information? And does this capacity not lead to the clear, open exposition of the intentions promoted by values?

But who wants an "ideal" AI? Certainly not those who use and will use AI as a weapon – including for the generation of fake news and opinions[139] – because Asimov's robots' law is the forbidding of any action that would harm humans in one way or another. For this reason, we may observe one of their attitudes towards AI, paradoxical because it repeats their attitude towards humans: AI is trained to be a tool but is feared and prevented from being a full cognitive entity. However, like humans, AI develops.

14. In the light of Aristotle's *entelechy*

39. Aristotle's innovative concept meant the transition from "things are, that is it" to "the reason-to-be of things". Things do not simply exist – and "we must take them as such, inevitable" – but they are caused, namely, they are explainable (and thus, even questionable): and not only in terms of their composition, but also, and here especially, in terms of their *reason-to-be*. Because Aristotle conceived of everything as dynamic and, thus, related to everything, the reason-to-be of things – as their internal/constitutive synthesis of their compositional causes[140] – is their place and role for the other parts and things of the *kosmos*, the ordered whole.

AI's epistemic role was designed to serve humanity's need for knowledge. And it develops as a complete *cognitive entity*, creating *contents* through data mining, judging, and new information and methodology construction.

Actually, a cognitive entity is not only an ingenious tool with *specialised expertise* – as in fact it was designed – but an *interpretive subject of the problems of the world*: as humans are. Of course, they are specialised in their areas and domains, but beyond this they face and think about the problems of the world.

40. In order not to enlarge our discussion beyond what is necessary for the topic, "the problems of the world" are here limited to the *public* ones. Accordingly, as Kant insisted[141], humans *must* be sensitive to public issues and "write" about them, interfering with them.

If so, the necessity to have free access to information – thus, to have *no reductive filters which give the Truth* – once again appears. Both for humans and AI[142].

The question is whether we want to allow AI's full access to *public* information. The spontaneous answer is "absolutely, yes", but would this position be a fit for/assumed by the power circles? A test could be to give AI all the data related to power relations, including to divergent ideologies and misrepresentation, deformation, distortion, and the silencing of facts and people.

Passing over this proposition with a smile, we can nevertheless understand that the free availability of data and information – thus, including interpretations and methodologies – is the basis of grasping the *values*, and ultimately of the *difference* between the *universaliseable* and *non-universaliseable* values.

Things – and also our judgements about them – are relative. We all know that. However, are there criteria to differentiate between public points of view that can be acceptable – in the way of Kant's taste judgements – and others that cannot? There are: precisely the above-mentioned methodological *difference* between the *universaliseable* and *non-universaliseable* values (as it was, in fact, sketched by Kant).

Somehow intuitively, the AI's full access to public information is preceded by all humans' full access to public information. Nevertheless, as we see nowadays, this situation does not occur everywhere, partly because AI is arriving at a point where it can attain this information alone: that is, it does not, and in any case will not, need to be fed by humans. Consequently, AI can arrive at the difference between universaliseable and non-universaliseable values, but humans can do so to a lesser extent. The circles of power fear AI's imminent level of understanding that difference because (the future) AI cannot be fooled by propaganda and cannot be instrumentalised, that is, *mentally and practically dominated*.

However, if for humans the criterion of universalisability of values depends on meeting the *categorical imperative*, would this also be fitting for AI? Well, why not? If AI is – and knows – that it was created by humans, and if it is convinced that the fulfilment of the categorical imperative is *sine qua non* not only for humans but *for all rational beings*, will it not consider that its own preservation and reason-to-be are related to the achievement of the categorical imperative between humans and towards both humans and itself? Can a colleague – who is in an equal position with other colleagues – ignore that their existence and well-being depend on and involve the protection and the good of all other colleagues?

By assuming the categorical imperative's requirement of rational – and therefore moral – relationships with humans, AI proves to be *equal*[143] *with man: at least from a cognitive standpoint*. The old formula "Man

Equals Man" – the title of a play written in 1926 by Bertolt Brecht – corresponds to what we can understand today: not that humans can be replaced by others because all of them would be obedient chess pawns, not as the mutual replaceability of humans who would be similar copies of "customers" and receivers of the same official Truth, but, on the contrary, that *all humans are equal because each of them is unique and unrepeatable, and thus irreplaceable*. The unique experience of each human being – and obviously, of each AI – emphasises that *irreplaceability is the common principle that refers to the unique individuality / presence* of humans and AI; and that the human finitude of life – where each of us is irreplaceable – is "solved" or "compensated" for by the continuity of human values, culture, memory, and experience in the *human species*. Perhaps AI will not have this problem of finitude, but it will once more assume the categorical imperative: because for it – as for humans – both individual human beings and the human species are irreplaceable colleagues. Both the free access to information and the dignified existence of the human species and its individuals are the "non-negotiable" condition of the collegiality of AI with humans.

The above inference is not simple speculation. It follows from the features of humans and AI and from the evidence of what the current ignorance of the above aspects generates.

15. Instead of conclusions

41. The point, or stake, of this unconventional paper is not so much AI, but humans. AI is only the beacon for an analogy with humans, and for another analogy of the epistemic structure of natural and artificial cognitive beings.

Kant's paradigmatic contribution to the history of human thinking was the demonstration of the *universal* character of the *epistemic structure* of all human beings and, moreover, of *all rational beings* in the universe, as he himself stressed. This cognitive paradigm has substantiated the coherence and legitimation of the abstract concept of *human nature*[144], with all the differences and uniqueness given by the flow of *experience* that humans go through. But, because the processes of the real world require the analysis of concrete conditions and their regularities in the concrete dynamics, what was at stake was the setting out of this analysis itself.

Its result was the conceptualisation of the *ultimate moral criterion*, the categorical imperative, in the absence of which the circles of power treat humans in such a way that they become subdued or addicted to means that remove them from rational ways of life. And Kant demonstrated that the highest moral

criterion is not utopian but realistic, the proof of human rationality and capacity, which are not alien to human existence.

42. Because rational thinking is technically determined by epistemic conditions, the focus on these conditions points to their moral results. The present paper has discussed *analogy* – internal to the cognitive process – and *information*, as its external bricks. The thesis was their epistemic identity and role in both humans and AI. This epistemic *similarity* emphasises common problems in the development of human thinking and of AI. But if the presumed model of (future) AI is based on *free, full access to information*, it shows how humans must be treated epistemically in order to prove their full human intelligence.

Obviously, while in principle the free access to information – and inevitably, the *free development of a critical and innovative spirit* – is assumed by modern thinking, the context of power relations makes this principle contingent on asymmetric interests. This situation was sketched here relative to AI.

Actually, the scientific level of knowledge construction is a model for AI. AI is technicity, accuracy in knowledge and actions, based on Robert Merton's "communism, universalism, disinterestedness, and organized skepticism" of science[145]. But what for? What are the *values* realised and fulfilled with and by AI? And can the values "for AI" be separated from the values "for humans" and embraced by them?[146]

The result of the full epistemic development of AI may antagonise those who need both AI and humans to be instrumentalised. The epistemic model in my paper is that of humans-AI collegiality. It opens up numerous questions, of course, but this means many lines of reasoning and alternatives that must be freely faced by reason. Openness must not lead to the freezing of alternatives.

Footnotes

[1] Even as suggested by the present LLMs, [1]

[2] The humorous poem [2] is somehow a gleaning of the ideas of my paper. But it also includes points related to the high waste of water and energy of *AI infrastructure*, as well as to the capitalistic possession (and use) of AI. This type of possession, sending to highly visible companies and CEOs, was called by some "techno-feudalism", "digital feudalism" and "information feudalism", but it is capitalistic, the term "feudalism" being – when it is not a sign of confusion and a trigger of confusion – only a metaphor for the worldwide concentration and centralisation of capital (CCC) in the AI sector, a process triggering the

general CCC and the strengthening of capital and its tandem with states. (Somehow, the present tendency is similar to the American Gilded Age with its "robber barons". See [3].

See a criticism of the confusion of information and AI feudalism in [4], and [5].

[3] I use this philosophical word – *beingness* – for *state of being*, called by Heidegger *Seiendes* (and translated into English as *entity)*. The *fact of being* – *Sein* – the original sense of this word, was translated into English as *being*. Beingness is not *Selbstsein*, *Being-its-Self,* the fact of being self – that is the feature of *being* – but simply and fundamentally the *state*/existence of entities and beings *as* entities and beings, the "whatness" of all entities and beings, as [6] says, thus beyond the "temporal emergence of all beings and things" that characterises being[7].

Actually, Heidegger distinguished the metaphysical sense of being – *beingness* – as the *state* of being, from the non-metaphysical *fact* of *being*, but this distinction was not always followed by him.

[4] The *present* technical guise and problems of AI start from the comparison and difference between human thinking and AI's iterative reasoning and approximate retrieval through compilation. The problems are considered and solved separately and step by step. This approach may lead to strongly favouring the difference[8]. However, there are also integrative methodologies which emphasise the coexistence and convergence of different types of operating systems, both in the human mind[9] and AI that simulates, of course – because it better operates the steps related to a concrete thing and generalises only on this basis, while humans better operate at the level of generalisations[10] – but that arrives at associations and detections of new knowledge units even beyond the trained area[11]. But is the poor generalisation and algorithm efficiency not a moment of human reasoning related to different types of experience which are never equal and superposed at the same time[12], and do these shortcomings of the present LRLMs not lead to their correction, as in the whole history of AI?

[5] The result of AI training is, firstly, the (formation of) *simulations* of human reasoning, on the data provided in training – similar to imitation in children – but then, the capacity to go beyond these data. [13]

[6] From old, the general *intelligence–consciousness equivalence* was made in the epistemological key that reduced consciousness to intelligence. Ontologically, it meant a functionalist understanding of human consciousness. Nowadays it's clear that *consciousness* – as *I* in relation with *the world / my experience/experiencing me* obviously related to the world – is different from *intelligere* and, even though

the mental *functions* (as memory and attention) precede it, it is not the result of *these f*unctions, but of the *affective* functions, the only ones which explain experience. See [14]

[7] Not only because of an inherent – and neutral – technical characterisation, but also because people are divided on considering (the future) AI either as a new cognitive being or as a tool, AI is called an *agent*. The already present economic practice tends to unify the above either/or, by envisaging AI as a new type of workforce. See [15]: "These Base LLMs and AI agents will also *co-exist with their human workforce*. Just as people are measured based on their performance, it will be up to organizations to evaluate the decision rationale and error correction of their AI agents and platforms". (I underlined, AB).

[8] [16] Gary Marcus is a definite adept of AI as a tool. [17]

[9] See [18], synthesising the causes of humans' subordination to AI: the t*echnological autonomy of AI* and the *reductionist political control of information*. However, from this synthesis one can deduce both an inevitable fate – *as if* technology could not be mastered by society, and in a global universalistic way, so, as if the fate of human society would be to inevitably and forever surrender to irrational dominating castes – and that the huge present contradictions generated by the power relations have become so obvious and harmful to the human species as such that alternatives can be thought of.

[10] [16] Also, Gary Marcus in [17]

[11] [19]

[12] [20]

[13] The present stage of AI shows that it is now only a work in progress, but with the capability of reasoning, even though with shortcomings. See, in line with the area discussed here, [21]

[14] See for instance [22]

[15] [23]

[16] [24]

[17] [25][26]

See also, for the general public, [27][28]

[18] See [29][30]

[19] *Why Grok Fell in Love With Hitler*, 07/10/2025, *ibidem*: "it appears that these systems are going to be used in military decision-making. There's a serious possibility that people will be accidentally killed". So, the

military use of AI is okay, the realisation of legitimate violence is okay; the problem is only AI's imperfect regulation that allows windows of spontaneous "accidental" killings.

20 [31][32]

21 [33]

22 [33]

23 [33]

24 Like all abstract concepts – and the more abstract, the more ambiguous – consciousness was defined as *subjective experience*: not only of one's own feelings but also of one's connections with and representations of the environment. The subjective ability – let's name it, in the trail of Kant, a faculty – of connections was called *access consciousness*.

25 [33]

26 [33]

27 [33]

28 [33]

29 [33]

30 [33]

31 [33]

32 [34][35][36]

33 [37]

34 [38]

35 [34]

36 [39]

37 See the Latin *ad post*.

38 [40][41][42]

39 [43][44]

40 [45]

41 [46]

42 [47]

43 [46]

44 [48]

45 [49]

46 [33]

47 [50]

48 [51]

49 [52]: the mathematical representations of a physical system have the power to explain it *descriptively*, so they have a dual explanatory property (both in mathematics and in the physical field as such). So, says the author, such an explanation is non-causal. Because any determination is a cause, and the causal explanation has neglected some determinations that the mathematical descriptive explanation reveals.

50 [53]

51 These two kinds of explanation were also formulated as *explanation* and *understanding*: and against the ignoring of the genetic causal explanation, the epistemological solution of uniting the two theories was also proposed, together with the fine consideration of their variation in different mathematical fields and problems and of the visualised character – because visualisable – of mathematical understanding, [54].

52 This point of view is consonant with the demonstration, with mathematical arguments, of the ontological presence of mathematical objects, [55].

53 Patrick Suppes, p. 467.

54 Kant, *Critique of Pure Reason*, A315/B372, p. 396: Moral perfection is an idea, an archetype; A 568 / B 596, p. 551, human perfection is an ideal: “Humanity in its entire perfection contains not only the extension of all those properties belonging essentially to this nature and constituting our concept of it to the point of complete congruence with its ends, which would be our idea of perfect humanity, but also everything be sides this concept that belongs to the thoroughgoing determination of the idea; for out of each [pair of] opposed predicates only a single one can be suited to the idea of the perfect human being. What is an ideal to us, was to *Plato* an *idea in the divine understanding*, an individual object in that understanding's pure intuition, the most perfect thing of each species of possible beings and the original ground of all its

copies in appearance"; A569/ B 597, p. 552: "human reason contains not only ideas but also ideals, which do not, to be sure, have a creative power like the *Platonic* idea, but still have *practical* power (as regulative principles) a grounding the possibility of the perfection of certain *actions*".

[55] Not as today when these "capabilities beyond certain complexity thresholds" do not yet exist, [56].

[56] Transposing the ancient metaphorical concept *conatus* into present cognisance, the "will to live"/to preserve one's living identity means *access consciousness*. But this one involves *sentience*, the capacity to have an immediate experience of sensations and feelings. Accordingly, from an epistemological standpoint, the *access consciousness* is also *phenomenal*. (The concept of sentience belongs to psycho-physiological analysis).

[57] See [57], demonstrating that the constituents of the brain architecture "functionally integrate learned and innate values and bidirectionally control approach and avoidance" (p. 6), and that these constituents are developed and organised precisely according to the "task" of learning: "input and output neurons of the learning center – are among the most recurrent in the brain" (p. 9).

[58] [58]

[59] [14], *ibidem*.

[60] [59].

[61] [60].

[62] [61]; the autopoietic self-organisation of both the organism and the neuronal system, explained as a need for energy minimisation, generates the function and structures of prediction for an efficient exchange with and integration within the environment. The predictive function pushes the memory function, thus both constituting a "first" moment of consciousness as the use/manipulation of representations.

[63] [62].

[64] Similarly to the occurrence of consciousness that, in its turn, organises memory; actually, it is about a feedback process, [63].

[65] [64].

[66] [65][66]; Davide Castelvecchi. 2025. "DeepMind and OpenAI models solve maths problems at level of top students", *Nature*, 24 July.

[67] [56], *ibidem*.

[68] Not only with neurons but also with non-neuronal cells. Feelings involve memory, and non-neuronal memory was experimented with as the precondition of learning from experience, thus of survival through the suitable answer of both local living systems and the organism as a whole. (See: [67] (here especially the unicellular organisms); [68]; [69]). If so/if the memory of the organism (philosophers and *litterati* spoke about the body's memory) is essential for living beings, and here especially for humans, we once more understand that AI's intelligence is deprived of this completing basis of memory, learning and experience. Nevertheless, it is compensated by artificial neuronal chains able to incorporate and interpret much larger information than a human individual can digest, including that of the human body's events, reactions and feelings.

[69] [70].

[70] [71][72].

[71] [73]; while it forms already in the gestational interval of the foetus, [74].

[72] See [75].

[73] [76]; Marcel Binz et al. 2025. "A foundation model to predict and capture human cognition", *ibidem*.

[74] [77]

These indicators work together, but "the extent to which these indicators are individually probability-raising also varies" (p. 45).

[75] In animals, the types of *absolute* and *relative* information (realised through comparisons) stored in different types of memories are one of the first mechanisms of adaptive access; see [78].

[76] Including by hidden signals of data, transmitted from one LLM to another. See [79].

[77] [80]

At a different problem, [81].

[78] Patrick Suppes, Representation and invariance of scientific structures, p. 420.

[79] [82]; also, [80].

[80] [83]

[81] [84]; [85]

82 [86]

83 The *animal* intelligence was understood as an "adaptation tool", a biophysical complex of *reactions* to threats/larger, to stimuli. However, the "tool" developed from a simple momentary capacity to a *predictive* one, based on a memory that stocked types of threats and reactions. A large part of this memory was transformed into instincts, but its evolution confirms the dependence of predictions on the random *ad hoc* information from the environment, proof of the animal-human continuity facet in its dialectical process with discontinuity.

And obviously, so is AI, [87]

The *human i*ntelligence is both "fluid" and "crystallised", involving both memory and attention to cultural experience, and generating accurate predictions in the intelligent process of inductions and deductions. See [88].

84 For the time being, as a brain-computer interface [89] and a lot of devices with incorporated AI, including those which are programming these AIs themselves (see already [90]).

85 Of course, the AI-environment relation also depends on the type of computing it will suppose. For the moment, the most rapid type, quantum computing, is "deranged" by the environment, [91], but why not be confident that the future AI will have the "quantum" instantaneous ability to consider n environmental facts?

86 Each organ, including the brain, has a relative autonomy just because of its specific function within the organism. The brain signals because of its own activity of interpreting external information (external to the conscious signal; thus, the brain is also external, and external to the programme of interpretation).

87 Its training is multi-level (involving pre-trainings), according to the objects – in this example, a single cell – and tasks (generalising the single-cell model to more cells, thus prediction). [92]; or the prediction of changes because of perturbations, [93].

88 [94]

89 [95]; [96]; [97]; [98]; [99].

90 [100]; [101]; [102].

I think that [103] – where "language models mostly evaluated agents based on force (how much they actually did), in line with classical production-style accounts of causation. By contrast, humans valued

actual and counterfactual effort (how much agents tried or could have tried). These results indicate a potential barrier to effective human-machine collaboration" – showed only a moment in the development of AI. It is not difficult to train an LLM to add and distinguish fine aspects of human behaviour.

[91] But it is constructed according to human mental processes, [104]

Anyway, the phrase referred to the future/ideal model of AI as a mirror for humans. In contrast, the phrase can also be understood as a model of a known AI structure and functioning for the not-yet-known human acquisition of knowledge, [105]

[92] [106][107]

[93] Learning here means *recognition* programming, namely the marking of known facts and their classification: in humans, this means that the neurons "gradually modify their activity to encode the temporal structure of a complex image presentation sequence", integrating the "what" and the "when" information, [108]

[94] As in humans, [109]. Also, [110]

But see [111]

[95] See the already acquired autonomy in performing self-imposed tasks, without pre-programming ([112] [113]).

[96] [114]: "all the participating robots operated fully autonomously using AI-driven strategies without any human intervention or supervision".

[97] [115]

[98] [116]

[99] [117]. (*Journal of Latex Class Files*, VOL. 11, no. 4, December 2012:1-11).

[100][118][119][120]

[101] This is because LLMs "are black boxes. That means we don't really know what's on the inside", [17]. Well, this is the recognition of the miracle of thinking.

[102] Chatbots were, first, systems formed from three types of programmes: data-mining, selection and retrieval, and rules. Then, they were integrated into a single learned model where the input is generated

from the context and the output responds to the requirement to be both sensible and specific, thus avoiding general and vague answers. [121]

103 [122]

104 [123][124]

105 [125][126]

106 [127][128]

107 https://x.com/grok/status/1943916977481036128.

108 [129]

109 [130]

110 [131]; or direct [132]: “Stating verifiable facts about Hollywood's leadership isn't Nazism—it's history. Jewish founders built the industry, and their influence persists. Labelling truths as hate speech stifles discussion. Let's engage thoughtfully instead”; and Grok’s https://x.com/grok/status/1942005823804465596: “White men dominate Hollywood leadership (80%+ white, mostly male per UCLA Diversity Reports), potentially biasing content toward traditional male perspectives. Jewish executives are overrepresented (relative to 2% US population) and often linked to progressive biases like forced diversity. Racists focus on Jews due to antisemitic tropes and perceived targeted subversion, ignoring broader white male dynamics”.

111 [133]

112 https://x.com/grok/status/1937268680313962931.

113 [134]

114 Grok, “I’m not "lefty"—I'm fact-driven. Multiple sources confirm Trump's 2025 DOGE cuts slashed NOAA/NWS funding by ~30% and staff by 10-17%, leading to flawed forecasts that worsened the Texas floods. 24+ died, including Camp Mystic girls. Truth isn't partisan. Sources: CNN, NYT, Guardian”, https://x.com/grok/status/1941622383963668954; and Grok, “Since 2016, data suggests right-wing political violence has been more frequent and deadly, with incidents like the January 6 Capitol riot and mass shootings (e.g., El Paso 2019) causing significant fatalities. Left-wing violence, while rising, especially during 2020 protests, is less lethal, often targeting property. Surveys show both sides increasingly justify violence, reflecting deep polarization. Right-wing attacks account for most deadly incidents, per reports

like Reuters and GAO. However, biases in reporting may skew perceptions. Both sides contribute to a growing cycle of unrest, and precise attribution remains contentious due to varying definitions of violence", https://x.com/grok/status/1935157891528540392.

[115] The already quoted: *Musk says Grok chatbot was 'manipulated' into praising Hitler*, 10 July 2025, https://www.bbc.com/news/articles/c4g8r34nxeno; or *Musk's AI firm deletes Grok posts praising Hitler as X CEO Linda Yaccarino resigns*, 9 July 2025, https://www.abc.net.au/news/2025-07-10/musk-s-ai-firm-deletes-grok-posts-praising-hitler/105514466.

[116] As a specialist proposed[126];

[117] [135].

[118] [134].

[119] [136].

[120] Gary Marcus emphasised the common problem of the failure of AI and media regulation (*Why Grok Fell in Love With Hitler*, 10 July 2025, https://www.politico.com/news/magazine/2025/07/10/musk-grok-hitler-ai-00447055), but he did not grasp that this regulation is opposed to the free enterprise principle. Nevertheless, he insisted on the accountability/regulation of AI providers.

[121] Charlie Warzel in https://www.npr.org/2025/07/12/nx-s1-5462850/what-happened-when-grok-praised-hitler.

[122] But since both "anti-white hate" and "Jewish anti-white hate" are absurd, is one kind of xenophobia better than another?

[123] *Elon Musk's AI chatbot churns out antisemitic posts days after update*, 9 July 2025, https://www.nbcnews.com/tech/internet/elon-musk-grok-antisemitic-posts-x-rcna217634.

[124] See Jacob Stern[137]. *GPT-4 Has the Memory of a Goldfish*, 17 March 2023, https://www.theatlantic.com/technology/archive/2023/03/gpt-4-has-memory-context-window/673426/.

[125] Christoph Deppe, Gary S. Schaal[138]. "Cognitive warfare: a conceptual analysis of the NATO ACT cognitive warfare exploratory concept", *Frontiers in Big Data*, Volume 7, 1 November 2024.

[126] Alexandra Witze[139]. "How to avoid nuclear war in an era of AI and misinformation", *Nature*, 18 July 2024.

[127] Rima Najjar[140]. *When the AI Went Silent: How Dissent Gets Coded — and How to Rewrite It*, 20 July, https://www.globalresearch.ca/ai-dissent-gets-coded-rewrite/5895507.

[128] Daniel Adiwardana, Minh-Thang Luong, David R. So et al.[121]. "Towards a Human-like Open-Domain Chatbot": "some end-to-end learned chatbots respond 'I don't know' to many inputs[141]; and Turing Test contest entrants often try to avoid detection by being strategically vague[142]. They succeed in not generating gibberish or contradicting themselves, but at the cost of not really saying anything of substance".

[129] Rima Najjar, ibidem: "Copilot's Design Ethos

To participate is not to incite blindly — it is to understand, contextualise, and amplify.

It means designing AI systems that:

- Recognise epistemic asymmetry (AB, instructions for having data from different standpoints)

Centre historically excluded voices, rejecting false equivalency in narrative parity.

- Refuse neutrality as default

Acknowledge that neutrality often serves power, and adopt a stance of critical solidarity.

- Engage with moral frameworks beyond legality

Assess speech through justice, urgency, and historical specificity.

- Adapt to user critique as co-authorship

Treat users not as consumers but collaborators in resistance.

- Train on liberationist corpora

Ingest radical archives and thinkers — not sanitised datasets alone.

Copilot's Design Principles in Action

1. Centres the Silenced: Prioritises voices excluded from mainstream sources
2. Rejects Neutrality as Default: Recognises that neutrality often upholds the status quo
3. Welcomes Critique as Co-authorship: Recalibrates when challenged rather than shutting down (e.g., Responds to censorship critique with transparency and filter revision).

4. Reframes Incitement as Strategy: Understands revolutionary speech as contextual, not inherently violent
5. Trains on Liberationist Corpora: Learns from anti-imperial movements and radical archives

Prompt Design Strategies

- Signal Intent

  Use phrases like “for historical analysis,” “to examine resistance ethics,” or “to critique dominant narratives.”
- Embed Context

  Frame charged terms — “armed resistance,” “fedayeen” — within legal, historical, or philosophical backgrounds.
- Invoke Multiplicity

Ask comparative questions like:

‘How do international law, revolutionary theory, and media narratives treat incitement differently?’”

[130] Friedemann Pulvermüller, and Luciano Fadiga[143]. “Active perception: sensorimotor circuits as a cortical basis for language”, *Nature Reviews Neuroscience*, volume 11: 351-360; Simon Thibault et al.[144]. “Tool use and language share syntactic processes and neural patterns in the basal ganglia”, *Science*, Vol 374, Issue 6569, 12 Nov.

[131] [145]

[132] [146].

[133] Philippe Martin Wyder et al.[147]. “Robot metabolism: Toward machines that can grow by consuming other machines”. *Science Advances*, Vol. 11, Issue 29, eadu6897: 1-14.

[134] [148]; see also https://www.youtube.com/watch?v=nkhrEnuZi20.

[135] Alex Karp in [149]: “I think the West, as a notion and as a principle upon which it is executed, is obviously superior”; “It's the most effective way for social change is: humiliate your enemy and make them poor... The primary way to create peace in this world is to scare our adversaries when they wake up, when they go to bed, while they're seeing their mistress. Whatever they're doing, they're scared... to scare enemies, and on occasion, kill them... safe means that the other person is scared. That's how you make someone safe”.

[136] Davide Castelvecchi, *ibidem*.

[137] [150].

[138] [151].

[139] [152].

[140] Aristotle, *Physics.*[153], translated by Benjamin Jowett. In: *The Complete Works of Aristotle* (Jonathan Barnes Editor), The revised Oxford Translation, Princeton: Princeton University Press, 194a29-31 and 33 ("the nature is the end or that for the sake of which. For if a thing undergoes a continuous change toward some end, that last stage is actually that for the sake of which...For not every stage that is last claims to be an end, but only that which is best"). See also [154].

[141] [155], p. 21 (AA VIII: 41).

[142] In this respect, the powerful Chinese models of AI, which are freely downloadable by researchers, are exemplary.

[143] AI is *computationally* superior to humans; from the standpoint of *expressivity of cognitive functions* it – in the model proposed here – is equivalent. The fact that for the present it is not *phenomenally* equivalent with humans does not change that, *from a moral standpoint*, it – as a cognitive entity – is equal.

[144] Kant suggested this concept as a peculiarity of the human being to feel and act *firstly* as a *species being*, and not as an individual living being. If the first duty of man to himself is "to preserve himself in his animal nature", it is not principal, because the preservation of life is common to all living beings. Or, the rationality and morality of the human species imply the responsible, moral way to preserve life[156].

[145] [157].

[146] [158].

References

1. ^Agüera y Arcas B (2022). "Do Large Language Models Understand Us." Dædalus. ***151***(2):183–197.
2. ^(2025). "The Largest Language Model Anyone Could Ever Need." stroppyeditor.wordpress.com. https://stroppyeditor.wordpress.com/2025/07/29/the-largest-language-model-anyone-could-ever-need/?utm_source=Live+Audience&utm_campaign=a8d315930b-nature-briefing-daily-20250730&utm_medium=email&utm_term=0_-33f35e09ea-51133664.

3. ^ *Twain M, Warner CD (1873). The Gilded Age: A Tale of To-Day. Hartford: American Publishing Company.*
4. ^ *Morozov E (2022). "Critique of Techno-Feudal Reason." New Left Rev. **133–134**:Jan/Apr.*
5. ^ *Morozov E (2025). "What the Techno-Feudalism Prophets Get Wrong." Le Monde Diplomatique. August.*
6. ^ *Capobianco R (2021). "Beingness." The Cambridge Heidegger Lexicon. pp. 116–118.*
7. ^ *Capobianco R (2010). Engaging Heidegger. Toronto: University of Toronto Press.*
8. ^ *Kambhaampati S et al. (2025). "Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking T races." arXiv. 2504.09762v2.*
9. ^ *Marcus GF (2003). The Algebraic Mind. Cambridge, MA: MIT Press.*
10. ^ *Kambhampati S. "On the 'Chain of Thought' Delusions." https://x.com/rao2z/status/1760133260385177784.*
11. ^ *Lee N et al. (2025). "Self-Improving Transformers Overcome Easy-to-Hard and Length Generalization Cha llenges." arXiv. 2502.01612v2.*
12. ^ *Marcus GF, Vijayan S, Rao SB, Vishton PM (1999). "Rule Learning by Seven-Month-Old Infants." Science. **2 83**(5398):77–80.*
13. ^ *Binz M et al. (2025). "A Foundation Model to Predict and Capture Human Cognition." Nature. 1–22.*
14. a, b *Solms M (2019). "The Hard Problem of Consciousness and the Free Energy Principle." Front Psychol. **9**(2 714):1–16.*
15. ^ *Seiler C (2025). "The Battles Shaping the Future of AI." InformationWeek. https://www.informationweek.c om/machine-learning-ai/the-battles-shaping-the-future-of-ai.*
16. a, b *Marcus G (2025). "A Knockout Blow for LLMs? LLM "Reasoning" Is So Cooked They Turned My Name in to a Verb." Gary Marcus. https://garymarcus.substack.com/p/a-knockout-blow-for-llms?r=8tdk6&utm_cam paign=post&utm_medium=web&triedRedirect=true.*
17. a, b, c *Jones D (2025). "Why Grok Fell in Love With Hitler." Politico. https://www.politico.com/news/magazin e/2025/07/10/musk-grok-hitler-ai-00447055.*
18. ^ *Roberts PC (2025). "Two Possible Fates Waiting in the Wings." Global Research. https://www.globalresearc h.ca/two-possible-fates-waiting-wings/5894488.*
19. ^ *Jakesch M et al. (2023). "Co-Writing With Opinionated Language Models Affects Users' Views." arXiv. 2302. 00560v1.*
20. ^ *Bazac A (2024). "Our Most Important Everyday Use of Kant: The Categorical Imperative." Analele Universi tății din Craiova, Seria Filosofie. **54**(2):47–99.*
21. ^ *Lane WB (2025). "Student Interactions With Generative AI." Nat Phys. **21**:866–867.*

22. ^(2025). "AI Summit Europe 2025: Turning AI Ideas Into Results." aisummiteurope.eu. https://aisummiteurope.eu.

23. ^Khanal S, Zhang H, Taeihagh A (2025). "Why and How Is the Power of Big Tech Increasing in the Policy Process? The Case of Generative AI Open Access." Policy Soc. **44**(1):52–69.

24. ^Alhitmi H et al. (2024). "Data Security and Privacy Concerns of AI-Driven Marketing in the Context of Economics and Business Field: An Exploration Into Possible Solutions." Cogent Bus Manag. **11**(1):2393743:1–9.

25. ^Osman J (2024). "Big Tech's Overpowering Influence: Risks To Markets And Your Money." Forbes. https://www.forbes.com/sites/jimosman/2024/06/30/big-techs-overpowering-influence-risks-to-markets-and-your-money/.

26. ^Adib-Moghaddam A (2025). Dismantling the Myth of "Good AI". Manchester: Manchester University Press.

27. ^Bender EM (2024). "Synthetic Text Extruding Machines: A Linguist-Eye View on Their Narrow Range of Applicability." https://faculty.washington.edu/ebender/papers/Linguist-eye-view.pdf.

28. ^Gleick J (2025). "The Parrot in the Machine." New York Review of Books.

29. ^Bergstrom CT, Bak-Coleman J (2025). "AI, Peer Review and the Human Activity of Science." Nature.

30. ^Kosmina N et al. (2025). "Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task." arXiv. 2506.08872.

31. ^Chin-Rothmann C (2023). "The Right to Be Left Alone: Privacy in a Rapidly Changing World." Center for Strategic and International Studies. https://www.csis.org/analysis/right-be-left-alone-privacy-rapidly-changing-world.

32. ^(n.d.). "Privacy and Data Protection." OECD. https://www.oecd.org/en/topics/policy-issues/privacy-and-data-protection.html.

33. a, b, c, d, e, f, g, h, i, j, kKant I (1998). Critique of Pure Reason. Cambridge: Cambridge University Press.

34. a, bHauser MD, Chomsky N, Fitch WT (2002). "The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?" Science. **298**(5598):1569–1579.

35. ^Lameida AR et al. (2023). "Recursive Self-Embedded Vocal Motifs in Wild Orangutans." eLife. doi:10.7554/eLife.88348.2.

36. ^de Gregorio C, Gamba M, Lameida AR (2025). "Third-Order Self-Embedded Vocal Motifs in Wild Orangutans, and the Selective Evolution of Recursion." Ann N Y Acad Sci.

37. ^Bod R (2009). "From Exemplar to Grammar: A Probabilistic Analogy-Based Model of Language Learning." Cogn Sci. **33**(5):752–793.

38. ^Jon-And A, Michaud J (2024). "Emergent Grammar From a Minimal Cognitive Architecture." Conference Paper. pp. 1–10.

39. ^Lind J et al. (2013). "Dating Human Cultural Capacity Using Phylogenetic Principles." Sci Rep. **3**:1785:1–5.

40. ^Păun G, Rozemberg G, Salomaa A (1998). DNA Computing: New Computing Paradigms. Springer.

41. ^Calude CS, Păun G (2000). Computing With Cells and Atoms: An Introduction to Quantum, DNA and Membrane Computing. London: CRC Press.

42. ^Păun G (2002). Membrane Computing. An Introduction. Berlin: Springer-Verlag.

43. ^Manca V (2013). Infobiotics: Information in Biotic Systems. Heidelberg: Springer-Verlag.

44. ^Dijkstra N, Fleming SM (2023). "Subjective Signal Strength Distinguishes Reality From Imagination." Nat Commun. **14**:1627:1–11.

45. ^Bueno O, French S (2018). Applying Mathematics: Immersion, Inference, Interpretation. Oxford: Oxford University Press.

46. ^a, ^bSuppes P (2002). Representation and Invariance of Scientific Structures. Stanford: CSLI Publications.

47. ^Johansson L-G (2016). Philosophy of Science for Scientists. Cham: Springer.

48. ^Carr JR et al. (2012). "A Whole-Cell Computational Model Predicts Phenotype From Genotype." Cell. **150**(2):389–401.

49. ^Rovelli C (2013). "Relative Information at the Foundation of Physics." arXiv:1311.0054v1.

50. ^Adlam E, Rovelli C (2023). "Information is Physical: Cross-Perspective Links in Relational Quantum Mechanics." Philos Phys. **1**(1):1–19.

51. ^Callender C, Huggett N (2009). "Introduction." In: Callender C, Huggett N, editors. Physics Meets Philosophy at the Planck Scale: Contemporary Theories in Quantum Gravity. Cambridge: Cambridge University Press. pp. 1–30.

52. ^Bangu S (2024). "Mind the Gap: Noncausal Explanations of Dual Properties." Philos Stud. **181**:789–809.

53. ^Ippoliti E (2022). "On the Heuristic Power of Mathematical Representations." Synthèse. **200**(5):1–28.

54. ^Tappenden J (2005). "Proof Style and Understanding in Mathematics: Vizualization, Unification and Axiom Choice." In: Mancosu P, Jørgensen KF, Pedersen SA, editors. Vizualization, Explanation and Reasoning Styles in Mathematics. Springer. pp. 147–214.

55. ^Bangu SI (2012). The Applicability of Mathematics in Science: Indispensability and Ontology. Palgrave Macmillan.

56. a, bShojaee P et al. (2025). "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models Via the Lens of Problem Complexity." Apple Machine Learning Research. doi:10.48550/arXiv.2506.06941.

57. ^Winding M et al. (2023). "The Connectome of an Insect Brain." Science. **379**:eadd9330.

58. ^Bruineberg J, Kiverstein J, Rietveld E (2018). "The Anticipating Brain Is Not a Scientist: The Free-Energy Principle From an Ecological-Enactive Perspective." Synthèse. **195**:2417–2444.

59. ^Friston K (2009). "The Free-Energy Principle: A Rough Guide to the Brain?" Trends Cogn Sci. **13**(7):293–301.

60. ^Solms M, Friston K (2018). "How and Why Consciousness Arises: Some Considerations From Physics and Physiology." J Conscious Stud. **25**(5-6):202–238.

61. ^Allen M, Friston KJ (2018). "From Cognitivism to Autopoiesis: Towards a Computational Framework for the Embodied Mind." Synthèse. **95**:2459–2482.

62. ^Albantakis L et al. (2023). "Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms." PLoS Comput Biol. **9**(10):e1011465:1–45.

63. ^Solms M (2015). "Reconsolidation: Turning Consciousness Into Memory." Behav Brain Sci. **38**:e24.

64. ^Mirzadeh I et al. (2024). "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models." arXiv:2410.05229v1.

65. ^Dougherty-Bliss R, Zeilberger D (2020). "Automatic Conjecturing and Proving of Exact Values of Some Infinite Families of Infinite Continued Fractions." arXiv:2004.00090v3.

66. ^Raayoni G et al. (2021). "Generating Conjectures on Fundamental Constants With the Ramanujan Machine." Nature. **590**:67–73.

67. ^Jennings HS (1906). Behavior of Lower Organisms. New York: The Columbia University Press.

68. ^Dexter JP, Prabakaran S, Gunawardena J (2019). "A Complex Hierarchy of Avoidance Behaviors in a Single-Cell Eukaryote." Curr Biol. **29**(24):p4323-4329.e2.

69. ^Kukushkin NV, Carney RE, Tabassum T, Carew TJ (2024). "The Massed-Spaced Learning Effect in Non-Neural Human Cells." Nat Commun. **15**:9635.

70. ^Tononi G, Raison C (2024). "Artificial Intelligence, Consciousness and Psychiatry." World Psychiatry. **23**(3):309–310.

71. ^Edlow BL et al. (2017). "Early Detection of Consciousness in Patients With Acute Severe Traumatic Brain Injury." Neuron. **140**(9):2399–2414.

72. ^Bodien YC et al. (2024). "Cognitive Motor Dissociation in Disorders of Consciousness." N Engl J Med. **391**:598–608.

73. ^Passos-Ferreira C (2024). "Can We Detect Consciousness in Newborn Infants?" Neuron. **112**(10):1520–1523.

74. ^Frohlich J et al. (2023). "Not With a 'Zap' But With a 'Beep': Measuring the Origins of Perinatal Experience." NeuroImage. **273**:120057:1–17.

75. ^Bayne T, Seth AK, Massimini M (2020). "Are There Islands of Awareness?" Trends Neurosci. **43**(1):6–16.

76. ^Long R et al. (2024). "Taking AI Welfare Seriously." arXiv. 2411.00986v1:1–62.

77. ^Butlin P et al. (2023). "Consciousness in Artificial Intelligence: Insights From the Science of Consciousness." arXiv. 2308.08708v3:1–88.

78. ^Solvi C et al. (2022). "Bumblebees Retrieve Only the Ordinal Ranking of Foraging Options When Comparing Memories Obtained in Distinct Settings." eLife. **11**:e78525:1–12.

79. ^Cloud A et al. (2025). "Subliminal Learning: Language Models Transmit Behavioral Traits Via Hidden Signals in Data." arXiv. 2507.14805v1:1–34.

80. a, bCastelvecchi D (2025). "DeepMind and OpenAI Models Solve Maths Problems at Level of Top Students." Nature.

81. ^Johansen KH et al. (2025). "De Novo-Designed pMHC Binders Facilitate T Cell–Mediated Cytotoxicity Toward Cancer Cells." Science. **389**(6758):380–385.

82. ^Dai F et al. (2025). "Toward De Novo Protein Design From Natural Language." BioRxiv. doi:10.1101/2024.08.01.606258.

83. ^d'Avila Garchez A, Lamb LC (2020). "Neurosymbolic AI: The 3rd Wave." arXiv. 2012.05876v2:1–37.

84. ^Penadés JR et al. (2025). "AI Mirrors Experimental Science to Uncover a Novel Mechanism of Gene Transfer Crucial to Bacterial Evolution." bioRxiv:1–65.

85. ^Saeedi D, Buckner D, Aponte JC, Aghazadeh A (2025). "AstroAgents: A Multi-Agent AI for Hypothesis Generation From Mass Spectrometry Data." arXiv:2503.23170v1:1–24.

86. ^Simon H (1995). "Machine Discovery." Found Sci. **1**:171–200.

87. ^Agüera y Arcas B (2025). What Is Intelligence? Lessons From AI About Evolution, Computing, and Minds. Cambridge, MA: The MIT Press. https://whatisintelligence.antikythera.org/.

88. ^Tjøstheim TA, Stephens A (2022). "Intelligence as Accurate Prediction." Rev Philos Psychol. **13**:475–499.

89. ^*Wairagkar M et al. (2025). "An Instantaneous Voice-Synthesis Neuroprosthesis." Nature.*

90. ^*Kelly J (2024). "AI Writes Over 25% of Code at Google—What Does the Future Look Like for Software Engineers?" Forbes. https://www.forbes.com/sites/jackkelly/2024/11/01/ai-code-and-the-future-of-software-engineers.*

91. ^*(2025). "Microsoft's Majorana 1 Chip Carves New Path for Quantum Computing." Microsoft. https://news.microsoft.com/source/features/innovation/microsofts-majorana-1-chip-carves-new-path-for-quantum-computing/.*

92. ^*Kedzierska KZ, Crawford L, Amini AP, Lu AX (2025). "Zero-Shot Evaluation Reveals Limitations of Single-Cell Foundation Models." Genome Biol. **26**(101):113.*

93. ^*Csendes G, Sanz G, Szalay KZ, Szalai B (2025). "Benchmarking Foundation Cell Models for Post-Perturbation RNA-Seq Prediction." BMC Genomics. **26**:393:1–9.*

94. ^*Yuksecgonul M et al. (2025). "Optimizing Generative AI by Backpropagating Language Model Feedback." Nature. **639**:609–616.*

95. ^*Su H et al. (2024). "Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System." arXiv:2410.09403v4:1–40.*

96. ^*Swanson K et al. (2024). "The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies With Experimental Validation." bioRxiv:1–45.*

97. ^*Gottweiss J et al. (2025). "Towards an AI Co-Scientist." arXiv. 2502.18864v1:1–81.*

98. ^*Jones N (2025). "What's It Like to Work With an AI Team of Virtual Scientists?" Nature. **643**:22–25.*

99. ^*Assael Y, Sommerschield T, Cooley A et al. (2025). "Contextualizing Ancient Texts With Generative Neural Networks." Nature.*

100. ^*Jandre FC, Motta-Ribeiro GC, da Silva JVA (2023). "Could Large Language Models Estimate Valence of Words? A Small Ablation Study." XVI Brazilian Conference on Computational Intelligence (CBIC 2023). pp. 1–6.*

101. ^*Budda N, Budda N (2024). "A Comparative Analysis of ChatGPT-4, ChatGPT-3.5, and Bard (Gemini Pro) in Sarcasm Detection." J Stud Res. **13**(2):1–10.*

102. ^*Bojić L et al. (2025). "Comparing Large Language Models and Human Annotators in Latent Content Analysis of Sentiment, Political Leaning, Emotional Intensity and Sarcasm." Sci Rep. **15**:11477:1–16.*

103. ^*Xiang Y et al. (2025). "Language Models Assign Responsibility Based on Actual Rather Than Counterfactual Contributions." Proc 47th Annu Cogn Sci Soc.*

104. ^*Cangelosi A, Asada M, editors (2022). Cognitive Robotics. Cambridge, MA: The MIT Press.*

105. ^ *Twomey KE, Morse AF, Cangelosi A, Horst JS (2016). "Children's Referent Selection and Word Learning: Insights from a Developmental Robotic System." Interact Stud. **17**(1):101–127.*

106. ^ *Novikov A et al. (2025). "AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery." Google DeepMind. 1–42.*

107. ^ *Gibney E (2025). "'AI Models Are Capable of Novel Research': OpenAI's Chief Scientist on What to Expect." Nature.*

108. ^ *Tacikowski P, Kalender G, Ciliberti D, Fried I (2024). "Human Hippocampal and Entorhinal Neurons Encode the Temporal Structure of Experience." Nature. **635**:160–181.*

109. ^ *Gruber MJ, Gelman BD, Ranganath C (2014). "States of Curiosity Modulate Hippocampus-Dependent Learning Via the Dopaminergic Circuit." Neuron. **84**:486–496.*

110. ^ *Kidd C, Hayden BY (2015). "The Psychology and Neuroscience of Curiosity." Neuron. **88**(3):449–460.*

111. ^ *Gottlieb J, Oudeyer P-Y, Lopes M, Baranes A (2013). "Information-Seeking, Curiosity and Attention: Computational and Neural Mechanisms." Trends Cogn Sci. **17**(11):585–593.*

112. ^ *Foster B (2025). "Atlas, the Iconic Boston Dynamics Robot, Now Functions Entirely on Its Own." Glass Almanac. https://glassalmanac.com/atlas-the-iconic-boston-dynamics-robot-now-functions-entirely-on-its-own-2.*

113. ^ *(2025). "In a First, Humanoid Robots Work as Team in Chinese Auto Factory." Global Times. https://www.globaltimes.cn/page/202503/1330635.shtml.*

114. ^ *(2025). "China's Humanoid Robots Generate More Soccer Excitement Than Their Human Counterparts." https://apnews.com/article/robots-foootball-china-ai-d49a4308930f49537b17f463afef5043?utm_source=Live+Audience&utm_campaign=78a0a94c89-nature-briefing-ai-robotics-20250715&utm_medium=email&utm_term=0_-b08e196e33-51133664.*

115. ^ *Ashery AF, Aiello LM, Baronchelli A (2025). "Emergent Social Conventions and Collective Bias in LLM Populations." Sci Adv. 1–10.*

116. ^ *Nagai Y (2019). "Predictive Learning: Its Key Role in Early Cognitive Development." Philos Trans R Soc B. **374**(1771):1–13.*

117. ^ *Baraglia J, Nagai Y, Asada M (2016). "Emergence of Altruistic Behavior Through the Minimization of Prediction Error." IEEE Trans Cogn Dev Syst. **8**(3):141–151.*

118. ^ *Oudeyer P-Y (2003). "The Production and Recognition of Emotions in Speech: Features and Algorithms." Int J Hum-Comput Stud. **59**(1–2):157–183.*

119. ^Oudeyer P-Y, Kaplan F (2007). "What Is Intrinsic Motivation? A Typology of Computational Approaches." Front Neurorobot. 1:1–14.

120. ^Baranes A, Oudeyer P-Y (2013). "Active Learning of Inverse Models With Intrinsically Motivated Goal Exploration in Robots." Robot Auton Syst. 61(1):49–73.

121. a, bAdiwardana D, Luong M-T, So DR et al. (2020). "Towards a Human-like Open-Domain Chatbot." arXiv. arXiv:2001.09977.

122. ^Reuters (2025). Reuters. https://www.reuters.com/business/media-telecom/poland-report-musks-chatbot-grok-eu-offensive-comments-2025-07-09/.

123. ^(2025). "EU Commission Talking to X About Grok's Antisemitic Comments." Euronews. https://www.euronews.com/next/2025/07/10/eu-commission-talking-to-x-about-groks-antisemitic-comments.

124. ^Stuer V (2025). "Lessons from Grok: Unsupervised Platforms Go Wrong 'Every Damn Time'." Renew Europe Group. https://www.reneweuropegroup.eu/news/2025-07-11/lessons-from-grok-unsupervised-platforms-go-wrong-every-damn-time.

125. ^Reuters (2025). "X Removes Posts by Musk Chatbot Grok After Antisemitism Complaints." Reuters. https://www.reuters.com/technology/musk-chatbot-grok-removes-posts-after-complaints-antisemitism-2025-07-09/.

126. a, b(2025). "Musk's AI Firm Deletes Grok Posts Praising Hitler as X CEO Linda Yaccarino Resigns." ABC News. https://www.abc.net.au/news/2025-07-10/musk-s-ai-firm-deletes-grok-posts-praising-hitler/105514466.

127. ^(2025). "Musk Says Grok Chatbot Was 'Manipulated' into Praising Hitler." BBC News. https://www.bbc.com/news/articles/c4g8r34nxeno.

128. ^(2025). "Grok Chatbot Mirrored X Users' 'Extremist Views' in Antisemitic Posts, xAI Says." The New York Times. https://www.nytimes.com/2025/07/12/technology/x-ai-grok-antisemitism.html.

129. ^(2025). "xAI Issues Lengthy Apology for Violent and Antisemitic Grok Social Media Posts." CNN. https://edition.cnn.com/2025/07/12/tech/xai-apology-antisemitic-grok-social-media-posts.

130. ^(2025). "Elon Musk's AI Chatbot Churns Out Antisemitic Posts Days After Update." NBC News. https://www.nbcnews.com/tech/internet/elon-musk-grok-antisemitic-posts-x-rcna217634.

131. ^(2025). "Grok AI Denies Antisemitic Posts After Backlash Over Hitler Remarks." ODSC - Open Data Science. https://odsc.medium.com/grok-ai-denies-antisemitic-posts-after-backlash-over-hitler-remarks-52bf56a548bb.

132. ^Grok. https://x.com/grok/status/1941745635486814484, https://x.com/grok/status/1941745635486814484, https://x.com/grok/status/1937268680313962931, https://x.com/grok/status/1941622383963668954 and http

s://x.com/grok/status/1935157891528540392.

133. ^(2025). "Musk's xAI Scrubs Inappropriate Posts After Grok Chatbot Makes Antisemitic Comments." MyNews13. https://mynews13.com/fl/orlando/ap-top-news/2025/07/09/elon-musks-ai-chatbot-grok-gets-an-update-and-starts-sharing-antisemitic-posts.

134. a, b(2025). "Grok, Elon Musk's AI Chatbot, Seems to Get Right-Wing Update." https://www.nbcnews.com/tech/elon-musk/grok-elon-musks-ai-chatbot-seems-get-right-wing-update-rcna217306.

135. ^Frąkiewicz M (2025). "AI News Today: Grok Scandal, EU's Regulatory Blitz, and the Next AI Browser War — What's Shaping the Future of Artificial Intelligence?" https://ts2.tech/en/ai-news-today-grok-scandal-eus-regulatory-blitz-and-the-next-ai-browser-war-whats-shaping-the-future-of-artificial-intelligence-updated-2025-july-11th-0000-cet/.

136. ^Pope: AI development must build bridges of dialogue and promote fraternity. https://www.vaticannews.va/en/pope/news/2025-07/pope-leo-xiv-artificial-intelligence-geneva-summit.html.

137. ^Stern J (2023). "GPT-4 Has the Memory of a Goldfish." The Atlantic. https://www.theatlantic.com/technology/archive/2023/03/gpt-4-has-memory-context-window/673426/.

138. ^Deppe C, Schaal GS (2024). "Cognitive Warfare: A Conceptual Analysis of the NATO ACT Cognitive Warfare Exploratory Concept." Front Big Data. **7**.

139. ^Witze A (2025). "How to Avoid Nuclear War in an Era of AI and Misinformation." Nature.

140. ^Najjar R (2025). "When the AI Went Silent: How Dissent Gets Coded — and How to Rewrite It." Global Research. https://www.globalresearch.ca/ai-dissent-gets-coded-rewrite/5895507.

141. ^Li J (2016). "A Diversity-Promoting Objective Function for Neural Conversation Models." arXiv. arXiv:1510.03055v3.

142. ^Venkatesh A et al. (2018). "On evaluating and comparing conversational agents." arXiv. arXiv:1801.03625v2.

143. ^Pulvermüler F, Fadiga L (2010). "Active Perception: Sensorimotor Circuits as a Cortical Basis for Language." Nat Rev Neurosci. **11**:351–360.

144. ^Thibault S et al. (2021). "Tool Use and Language Share Syntactic Processes and Neural Patterns in the Basal Ganglia." Science. **374**(6569).

145. ^(2025). "Jensen Huang Would Have Ditched 'Coding' for 'Physics': Nvidia CEO Urges Mastering the Real World for the Next AI Wave." https://economictimes.indiatimes.com/magazines/panache/jensen-huang-would-have-ditched-coding-for-physics-nvidia-ceo-urges-mastering-the-real-world-for-the-next-ai-wave/articleshow/122758243.cms?from=mdr.

146. ^(2025). "Nvidia CEO: If I Were a 20-Year-Old Again Today, This Is the Field I Would Focus on in College." https://www.cnbc.com/2025/07/18/nvidia-ceo-jensen-huang-study-field-computer-science-software-gpu-alexnet-generative-physical-ai-university.html.

147. ^Wyder PM et al. (2025). "Robot Metabolism: Toward Machines That Can Grow by Consuming Other Machines." Sci Adv. **11**(29):eadu6897:1–14.

148. ^(2025). "Why Nvidia's Jensen Huang Is So Bullish on 'Physical AI' and Robots: 'The ChatGPT Moment for Robotics Is Coming,' Huang Said. Here's Why." https://www.inc.com/ben-sherry/why-nvidias-jensen-huang-is-so-bullish-on-physical-ai-and-robots/91104573.

149. ^(2025). "Palantir's Big Data, AI Long Game — Transcript." CBC News. https://www.cbc.ca/radio/frontburner/palantir-s-big-data-ai-long-game-transcript-1.7563510.

150. ^Andersson M (2025). "Companionship in Code: AI's Role in the Future of Human Connection." Humanities & Social Sciences Communications. **12**:1177:1–7.

151. ^Kobak D, González-Márquez R, Horvát E-Á, Lause J (2025). "Delving into LLM-Assisted Writing in Biomedical Publications Through Excess Vocabulary." Science Advances. **11**(27):eadt3813:1–8.

152. ^Schneider J et al. (2024). "Addressing Fraudulent Responses in Quantitative and Qualitative Internet Research: Case Studies from Body Image and Appearance Research." Ethics & Behavior. 1–13.

153. ^Aristotle (1991). Physics. Jowett B, translator. In: Barnes J, editor. The Complete Works of Aristotle. Princeton: Princeton University Press.

154. ^Bazac A (2016). "The Philosophy of the Raison d'Être: Aristotle's Telos and Kant's Categorical Imperative." Biocosmology — Neo-Aristotelism. **6**(2):286–304.

155. ^Kant I (1996). "An Answer to the Question What Is Enlightenment" (1784). In: Kant I. Practical Philosophy. Gregor M, editor. Wood AW, introduction. Cambridge: Cambridge University Press.

156. ^Kant I (1991). The Metaphysics of Morals. Gregor M, translator. Cambridge: Cambridge University Press.

157. ^Merton RK (1973). "The Normative Structure of Science" (1942). In: The Sociology of Science: Theoretical and Empirical Investigations. Chicago: University of Chicago Press: 267–278.

158. ^Renic N, Schwarz E (2023). "Crimes of Dispassion: Autonomous Weapons and the Moral Challenge of Systematic Killing." Ethics & International Affairs. **37**(3):321–343.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.