Mind, Biology, and Value Alignment:
Precis of The Prospect of a
Humanitarian Artificial Intelligence

Carlos Montemayor
Philosophy Department
San Francisco State University, USA

Abstract

This is a short sketch of some central ideas developed in my recent monograph The Prospect of a Humanitarian Artificial Intelligence, published by Bloomsbury, London 2023. The monograph is available open access at library.oapen.org/handle/20.500.12657/61934. It illuminates the development of AI by examining our drive to live a dignified life. It uses the notions of agency and attention to consider our pursuit of what is important. It shows how the best way to guarantee value alignment between humans and potentially intelligent machines is through attention routines that satisfy similar needs. Setting out a theoretical framework for AI, the book acknowledges its legal, moral, and political implications and takes into account how epistemic agency differs from moral agency. Insightful comparisons between human and animal intelligence clarify why adopting a need-based attention approach justifies a humanitarian framework. This is an urgent, timely argument for developing AI technologies based on international human rights agreements.

1. Intelligence and Consciousness Are Not the Same

Research on artificial intelligence (AI) is already impacting society and science. If artificial general intelligence (AGI), or human level intelligence, is developed, the consequences for humanity will be permanent and profound. Much optimism and pessimism has been expressed in academia and in the public sphere (“social media” in particular) about these developments, with considerable confusion about what exactly is at stake. Some engineers are attributing rationality, consciousness, and emotions to the systems they are creating in what constitutes an unprecedented psychologizing of technology by the very people that know and build the machinery the best. Other developers and leaders of industry proclaim that it is important not to confuse the public and maintain that these are
mere tools that will help us improve our lives by automatizing and optimizing a vast amount of labor. Even in this second scenario of automatized optimization, there is anxiety about how such automation will affect the lives of workers, the environment, the public sphere, and democracy. But what is already surprising is the amount of divergence one finds within the industry with respect to how to characterize this technology as either an autonomous intelligence or a mere tool (see Mindt and Montemayor 2020 for discussion).

One of the main goals of the book is to bring clarity to this discussion, from both a psychological and philosophical perspective. This clarification is made in the spirit of delineating the nature and limits of AI in a scientifically responsible way. The book argues that intelligence is not the same as consciousness. This thesis receives considerable evidential support, and it has critical implications for the possibilities of engineering AI. For example, developing highly intelligent AGI will not necessitate the type of consciousness humans and animals have. Although AGI may not require any kind of consciousness at all, such a highly intelligent agent will be missing essential ingredients of human and animal cognition. But this consequence is very helpful if one wants to engineer and reproduce intelligence, because consciousness is the irreproducible subjective perspective of an organism, which seems to be an impossible starting point for the purposes of engineering design.

The dissociation between intelligence and consciousness also generates risks because our cognition and trust rely at least partly on the conscious awareness of our emotions and biological needs (see Montemayor et al. 2021). The book introduces distinctions that clarify this issue. A highly intelligent AGI may become intensionally equivalent to human intelligence with respect to epistemic goals and needs, but it can only be extensionally equivalent with respect to the moral and aesthetic needs that fundamentally depend on our being conscious organisms. This means that AGI will have similar needs and epistemic motivation, but a very different emotional or empathic life, if it has any at all. However, by being extremely good at representing human emotions it can approximate emotional intelligence, without ever being fully equivalent. How exactly is intelligence different from consciousness? There is the obvious difference that intelligence is contrastive and public – we evaluate the intelligence of people and contrast their capacities in admissions to schools, job applications, and prizes. Consciousness is a deeply private and subjective matter and we never think of it in contrastive and public ways, with the exception of patients that may have only minimal kinds of awareness.

However, one can be a lot more specific here, particularly about the

---

1 Extensional equivalence is superficial, defined in terms of performance. Intensional equivalence is based not only on performance, but also on the similarity of the contents and motivations of genuinely autonomous agents.
psychological and epistemic details. The book proposes that attention, understood as the mental action that allows us to select information in order to use it for planning and problem-solving, is the fundamental cognitive capacity required for intelligence in humans and animals. Attention is mental action under our control, and it takes us from selected contents to a conclusion that satisfies epistemic goals. This proposal solves two problems. The first is that attention is essentially a kind of epistemic agency, which means that we have *agental control* over what we attend or do not attend to. For this reason, we are also responsible for our attention. This is fundamental to understand why only agents are genuinely intelligent: because they are *autonomously* in control of their attentive capacities. AGI would need to be attentive in this way, autonomously, for it to count as an intelligent agent and not a magnificently complicated tool, entirely under our control.

Second, similar to the way transformers and large language models operate, the process of mentally moving from inputs to outputs through attention control does not require explicit rule following or step by step argument structure. The analogy with transformers or deep neural networks, however, is not perfect because, although attention is a kind of control over the mental actions of humans and animals, contemporary AI systems lack autonomy and full control over what they are doing. They still depend fundamentally on human intervention for data curation and guidance at different stages of production. In addition, unlike attention in humans and animals, the contemporary AI systems that essentially depend on data curation are, by definition, not interacting either with the environment or with other agents. These are two essential requirements for attention in humans and animals: autonomous control and a cognitive anchoring that can only be provided by interacting with the environment, including other agents. A non-autonomous AI system will be extensionally equivalent to aspects of human and animal intelligence because it will arrive at similar results if prompted by similar inputs. But only autonomous AI agents, intrinsically motivated by the imperative of satisfying their own representational and cognitive needs, can count as intensionally equivalent to human or animal intelligence.

Attention grounds epistemic agency without consciousness, and this is quite important for the development of AI. If intelligence is the optimal solution to problems that one needs to solve in a timely fashion, then most of this work can be done by attention and our capacities for inference. In fact, our inferential capacities are best understood as part of our attention capacities to transition from specific contents to a conclusion, in a way that requires agent-level control without necessitating a specific kind of phenomenology or subjective awareness (Montemayor 2019). Human subjective awareness and its phenomenology accompany the use of our attention, but it is attention that gets things moving in our minds, as is
explored in detail below.

What is the contribution of consciousness to human cognition, and to intelligence in general? Framing this question in terms of what an autonomous agent needs to do in order to behave intelligently, what does it mean to say that consciousness is not necessary for intelligence? Attention solves problems that have to do with knowing the world, representing the environment and learning from it, autonomously and independently; what contents are or should be salient to the agent, and what the agent should be curious about. But these representational and rational needs are just one part of the larger set of needs that biological organisms have. Equally important are the needs that depend upon their emotions and vivid experiences. The satisfaction of these needs necessitates the involvement of our conscious awareness.

One way of thinking about this is in terms of value alignment. Animals align their mental actions through joint attention. A large part of how they are epistemically aligned depends on the uniformity of their cognitive and representational needs. Animals autonomously satisfy these cognitive needs through their attentive control – they are genuinely intelligent agents, and we share with them similar cognitive and representational needs. Animals are, in this respect, not only extensionally equivalent to us regarding some capacities, such as vision, audition, olfaction, and planning. They are also intentionally equivalent to us, because they are intrinsically motivated to satisfy their own needs to represent the environment and interact with it. By contrast, if we develop an AGI that performs or outperforms humans in most tasks, but which is detached from the environment, and without genuine autonomous cognitive needs and motivations, then this system, impressive as it will be, can only be extensionally equivalent to us.

Consciousness is an entirely different source of value alignment. Our experiences and feelings dignify us by giving us a perspective on the world that is unique to us. This is related to the notions of moral dignity and moral status, which differ considerably from mere epistemic agency and the general capacity for problem-solving. Consciousness connects us with the evolution of life through our emotional needs and experiences, rooted in biology. Our conscious awareness is how our biology familiarizes us with the organic world through our unique perspective. It also provides an empathic source of familiarity with the experiences of others, certainly humans, and also many animals.

Consciousness may not be necessary for an ant to pay attention to its surroundings. This is not to say that an ant is not at all conscious, but rather, that consciousness is not essential to the problems an ant needs to solve in order to navigate its environment. A dog has similar representational and navigational needs, but it also clearly has some degree of conscious awareness and a robust emotional life. This is why we experi-
ence the suffering of a dog in ways that are similar to our experiences of human pain, and also why we don’t have similarly strong empathic reactions when someone kills an ant. This empathic connection is through what it is like for us to feel pain, and not how we could represent and pay attention to bodily damage.

This is important because consciousness is not essentially a kind of representation that allows us to satisfy cognitive needs. Rather, it is a deep connection with our biology, which makes us feel in certain ways, and which familiarizes us with the world through vivid experiences. Using other terminology, which is more relevant for the topic of AI consciousness, attention is algorithmic in structure because it proceeds from input to output by selecting and contextualizing information. Consciousness, by contrast, is homeostatic, because it stays at equilibrium and does not halt or come to an end once a goal is accomplished.

There are various views about the metaphysics of consciousness that may complicate this story, but this essential point about the difference in information processing between attention and consciousness would still hold, regardless of these complications. On any view of consciousness, our biology is the seat of our conscious awareness because our life sustains it, and likewise, on any view of consciousness the value of conscious awareness goes beyond mere goal satisfaction and intelligent performance. So even those who have views about, for example, the cosmic fundamentality of consciousness or its dependence on neural machinery, can accommodate these claims.

My view is that consciousness is a deeply biological phenomenon and that this is part of the explanation of why it dignifies our mental lives through the familiarity it affords to our unique perspective on the world. This perspective involves a different kind of autonomy than that of attentive control. It is the autonomy of life and metabolically independent growth. It might depend on the properties of organic assembly that underlie development, providing a field not of algorithmic information, but of metabolic integration. Through our emotions and vitality, consciousness provides us with the autonomy of subjective experience.

In this sense, consciousness is not a kind of control, but rather, a kind of connection with our ancient evolutionary history. More developed forms of consciousness, like those that interface with attention and long-term memory and planning, also depend on this ancient connection. Consciousness is clearly informative because its contents are highly specific – take the case of color, for instance. But conscious information is not algorithmic or merely representational. It is a constant, deep, and dignifying connection that stays “always on” during our wakeful lives and in dreams. Consciousness is what ultimately dignifies our cognition through our emotional lives.

This summarizes the most important arguments and distinctions of
the book with respect to philosophy of mind and epistemology, which are applied to AI. The book uses these distinctions to discuss how AI could be understood in terms of collective agency, the political implications this might have for shaping the public sphere, generating and distributing knowledge, as well as generating and distributing political power. It concludes with the proposal that an international regulation of AI, based on the human rights framework, is the best way to avoid militarizing AI, biasing it with parochial state agendas, and to guarantee that human dignity is protected as a whole, not on the basis of commercial or political interests, but because of its intrinsic worth.

2. What Does AI Teach Us About Our Minds?

In the remainder of this precis, the emphasis is on alignment and attention, with a focus on more technical issues. The amount of resources and expertise that are invested daily to develop AGI constitutes one of the largest experiments about the mind and its capacities, at an industrial scale, in the history of humanity. From a scientific perspective, this is an unprecedented opportunity to investigate the nature of intelligence and mental skills. It is not ideal that industrial secrets and financial interests taint this effort. But it certainly is epistemically beneficial to have so much collective energy around what is, deeply, a fundamental scientific and philosophical question: what is intelligence, and can it be produced artificially? There are multiple risks associated with industrially produced intelligence, which the book addresses in detail. But the focus here is on what we have learned about our own minds during this dramatic development, particularly the last couple of decades, and how these lessons relate to the distinctions and themes of the book.

First, as many of its proponents emphasize, one of the most important lessons about the tremendous success of deep neural networks and reinforced learning is that it works at all. The fact that a brain-inspired approach to computation surpassed in power and success the previous way of writing explicit instructions through code is amazing. One would have at least expected that both styles of programing would be roughly equivalent in the best case scenario for neural networks. But this didn’t turn out to be the case. To see more specifically why this is so, some details are needed. A key detail is that the system can be trained to produce a restriction on the weights of the network that is equivalent to a program – a solution to a specific problem. The system improves this program as it scales in compute power and the size of the database. Human explicit commands become less efficient as the system scales up.

Second, there is considerable uniformity with respect to how these systems operate. For instance, the transformer architecture can solve a wide variety of problems without changing the basic way the model works.
The system, in other words, doesn’t need to be tailored specifically to the task by the programmers, task by task, instruction by instruction, or detail by detail. The generality of these systems is, therefore, more independent from human intervention than the standard universality we were familiar with, specifically the notion of a universal Turing machine. These systems are universal approximators to solutions and programs, without the need for specific instructions or commands.

The lack of programming, understood as coded commands and explicit instructions, is perhaps the most important characteristic of these models with respect to what philosophers of mind and epistemologists need to learn from them. The system produces an effective and optimal program, a constrained network and the efficient compression of information, based on a cost function. This program is structured by the weights of the network, and the program produced effectively computes the function it is asked to perform. This second style of computing is so powerful that it has been adopted in most areas of programming. It resembles much more how our own mind works, as explained below, and it is certainly inspired by biology from its inception. Using this new approach to programming is so radically different to previous programming approaches that Andrej Karpathy (2017) called it “Software 2.0”. As Karpathy writes:

> It turns out that a large portion of real-world problems have the property that it is significantly easier to collect the data (or more generally, identify a desirable behavior) than to explicitly write the program.

There are certainly lessons for philosophy and psychology here. Although our minds don’t need massive databases to learn and meta-learn (i.e., learn how to learn), our minds also learn, extrapolate, boost content, and develop without explicit instructions, certainly not instructions that we consciously or voluntarily follow. Animals learn quickly, represent their environment faster after being born than humans – we have the privilege of very long childhoods by comparison – and succeed at multiple tasks that demonstrate intelligence and knowledge without any explicit kind of linguistic rule-following or instructions. Some animals can move around in their environment with remarkable knowledge of the world and of other agents acting on it right after being born. Our new style of computing approximates animal and human intelligence by not depending on explicit instructions, and to that extent, its biologically-inspired analogy with the brain has succeeded as a scientific approach to programing.

But there are also lessons for AI developers from the biological and philosophical side of things. Engineers and computer scientists should consider more carefully the philosophical background of their discussions, models, and terminology. It is true that attention operates similarly to transformers, particularly because attention selects information from a
body of evidence or data, and moves from contents that are relevant for the solution to a problem to its conclusion, compressing the complexity of the original data set to a specific output. But a key distinctive aspect of attention in humans and animals is that it is a manifestation of their cognitive autonomy, their genuine kind of mental agency, which is in constant interaction with the environment. This contrast between contemporary AI models and human cognition is more striking than the fact that we do not need large data sets to learn. Humans and animals have cognitive autonomy and are in a dynamic informational exchange with their immediate environment and with other agents. AI models are by definition trained on sets of data and they never interact directly with the world in an autonomous way in order to represent what is salient to them. Thus, one lesson from philosophy is that AGI models must be even more inspired by realistic biological models if they are to become autonomous and genuinely intelligent at some point.

One way of making more specific the requirement that attention is a kind of agency, a cognitive capacity of an embodied animal embedded in its environment, is that attention does not just solve any random set of problems thrown at it, but rather, that it solves a series of problems that are required to satisfy the cognitive needs of the agent. This is the definition of epistemic autonomy proposed in the book, which also constrains and reduces the type and number of problems that must be solved by the agent.

Some of the needs an autonomous agent must satisfy include the need to represent the environment accurately, know how to navigate the world, determine what to learn, and how to speed up learning through proper curiosity and creativity. An autonomous agent must also communicate with certain intentions, and detect the communicative intentions of other agents for communication to succeed. Attention helps agents satisfy these needs autonomously, as part of their cognitive control over the contents their mind processes. Any genuine AGI will, by necessity, satisfy this kind of autonomy. Otherwise, the AGI should be considered an incredibly sophisticated tool, or form of automation. Attention limits the ways in which the specific cognitive needs of an agent are satisfied, autonomously and in direct contact with the environment. This is certainly an important lesson from epistemology for the development of AGI.

Based on the difference between consciousness and attention, the book proposes that the value alignment problem should be divided into two problems. There is the epistemic alignment problem, which is that agents need to align their epistemic values concerning truth-seeking, problem-solving, learning, and knowledge in order for them to cooperate efficiently. Attention is the main capacity humans and animals use in solving these problems. But alignment would not be guaranteed by the individual effort of each agent trying to align their attentive capacities. A robust kind
of uniformity among the cognitive and epistemic needs of agents is also necessary. This occurs by default in humans and animals by having similar cognitive architectures and, crucially, by sharing the same environment. As an illustration of these issues for AI systems, the epistemic alignment problem for contemporary AI models can be subdivided as follows.

Alignment 1: To train the system, data curation must be done in such a way as to ensure quality, adequate size, relevance with respect to the problem specification, and ground truth. For example, if the problem is driving, then most of the alignment is epistemic: how to represent the environment, including other drivers and vehicles, how to classify objects as the car moves on, how to represent situations, such as an elderly person crossing the street, and so on. Clearly, ethical and legal issues emerge as one aligns the system at this first stage, but this is just to illustrate how difficult it is to achieve epistemic alignment, independently of ethical considerations. An autonomous vehicle should have a minimal degree of independence as it solves these problems in real time, hopefully through some kind of unsupervised learning about how to adapt when embedded in real environments. But this requires a second level of alignment, where the supervision is not about the original data set but about the performance of the vehicle.

Alignment 2: When the system is deployed, there is a second stage of human supervision where there is human feedback from drivers, at different locations, with different cultural standards and habits of driving, as well as social contexts, such as posh neighborhoods or crowded intersections. A cynical, although not entirely inaccurate, way of understanding this kind of alignment is in terms of “pleasing the customer”. But even this requires non-trivial epistemic alignment: what counts as a posh neighbor and how exactly does it differ from a very populated area with problematic intersections? Answering this kind of questions is also part of achieving ground truth: correctly representing a crowded intersection, with all its intricacies, helps the car understand when it should assert itself in order to compete for space. The risks for bias here are epistemic, in a way that they affect ethical alignment as well: drivers from a certain ethnicity are systematically taken for drivers at posh neighborhoods, and so on.

This example shows that satisfying the representational and epistemic needs of an agent, for instance the need to determine what is factual, which includes knowing that something can be verified, is extremely important for other kinds of alignment, including ethical and aesthetic alignment, as well as legal compliance. Human drivers are aligned because they have similar cognitive needs. There are various needs drivers must satisfy, including protecting pedestrians and not injuring people, and both are needs that include ethical and legal standards. Hence the importance of the hierarchy of needs proposed in the book. Drivers clearly need to be
able to represent the road, see clearly what signs say, and so on, in order to be capable of driving at all. But once these representational needs are satisfied, other needs must be at the forefront, such as safety needs. Attention provides the flexible and heuristic machinery that allows us to satisfy our cognitive needs smoothly and in real time.

The AI industry, therefore, needs to talk more responsibly when using terms like “attention” or “learning” because the way humans solve alignment problems, such as alignments 1 and 2, is through their autonomous agency and mental control. The way alignments 1 and 2 are currently approached in AI development depends entirely on human supervision. Humans and animals align their values by reducing the space of possible data sets that might be relevant, as well as by constraining the space of possible solutions to a given problem. The needs agents must satisfy and their autonomous attentive control help reduce these possibilities, allowing them to rank these problems in terms of importance by compressing them to a small set at any given moment. This kind of autonomy for alignment, grounded on the needs of an agent that is embedded in an environment that demands the satisfaction of such needs, is the source of intelligence in humans and animals. If this level of autonomy is reached for alignments 1 and 2 in an AI agent, that would entail that the system is an AGI.

With respect to the impact of AI on society, the book argues that we shouldn’t “engineer” our way out of all conflict and alignment problems. That would be both tyrannical and dehumanizing. What we need is enough alignment to protect human dignity and enough freedom to allow for variation, which is also part of human dignity. The international human rights framework can help us do this. It offers a kind of meta-alignment that has the advantage of making states responsible for implementing an international framework for human dignity in the development and deployment of AI, rather than their own domestic and military agendas. This is a powerful reason to resist the idea that everything should be done at the engineering level. To solve global problems or threats to humanity we need human institutions that protect human dignity.

Let me conclude with another lesson for philosophy from what Karpatth calls Software 2.0. Inference plays a central role in epistemology. Rationality is defined in terms of our capacity to provide reasons for our thoughts and actions, and this is typically taken to require our conscious assent to the argument structure that leads to the conclusion, and to intellectually “see” that the premises entail the conclusion. Rationality requires seeing the inference and taking it for what it is, a reason for the conclusion. The book argues against this view. In particular, it proposes that inferential reasoning requires attention, and that attention to inferences need not, and in many cases should not, depend on our conscious
assent for inferences to be both rational and under our control. What I want to emphasize here is how important it is to update our epistemology, in order to fully appreciate the non-trivial conditions that mental action, including inferential reasoning, require.

In a logic or math class, going through an argument and the rules that govern its validity require conscious understanding and reasoning. This is true and it is not in dispute. The question is how can animals, infants, and adult humans reduce the space of possible data sets and potential solutions to problems mentioned above when it comes to drawing a conclusion in an inference? Does all inferential reasoning require careful and slow conscious reflection on rules and their validity? It has become clear in recent epistemology that it is inadequate to explain inference, and bias in general, as a kind of conscious intuition. What is needed is an account of optimal and fast control for mental action, and this is what attention provides. As one moves from one mental content to the next, what matters is that the motion of our mind arrives where it is supposed to arrive, namely at a solution to a problem or at a reduction of the available data sets. In the case of inference, this means that we quickly and efficiently draw the conclusion. If we fail at this basic task, no amount of deep conscious meditation will help us.

More specifically, when we mentally act in order to transition from entering into an inference to its conclusion, and out of it in order to stop inquiry, no amount of conscious enlightenment will help if one gets, for instance, simply stuck midway in one’s conscious reflection. Real, world-constrained mental action needs to “move on” and get out of the inference, without getting permanently stuck, even if enlightened by the phenomenology of truth preservation. The inference needs to be under the autonomous and dynamic control of an agent with specific needs. These needs are vast and require enormous amounts of cognitive resources. They include the need to assert, to open inquiry, to close inquiry, as well as the need to be selective and to doubt. The attentive control of agents, in satisfying these needs, is the source of their epistemic responsibility (Fairweather and Montemayor 2017).

The lesson is, then, that we need to update our epistemology, from the ancient conception of individual conscious assent to the realm of biologically inspired models, including animal models and those we are using now in AI development. Because of the rise of AI we may also need to update our epistemology to the realm of technologically determined social epistemology, where AGIs become the most powerful knowledge producers and, for this reason, should be considered collective epistemic agents, as the book argues. These updates will bring fresh air to our debates, and make philosophy more relevant to the world of intelligence design.

References


Received: 01 December 2023
Accepted: 06 December 2023
Invited by the editors