Acta Scientific Neurology (ASNE)

Volume 9 Issue 3 March 2026

Research Article

# NeuroAI, Paleoneurology, and the Alignment of Superintelligent AIs with Human Values. Our Future Depends on It.

**Margaret Boone Rappaport[1]* and Christopher J Corbally[2]**

[1]*Former President, Policy Research Methods, Inc., USA*

[2]*Vatican Observatory, University of Arizona, USA*

***Corresponding Author:** Margaret Boone Rappaport, Former President, Policy Research Methods, Inc., USA.

**DOI:** 10.31080/ASNE.2026.09.0896

**Received:** January 29, 2026

**Published:** February 28, 2026



## Abstract

**Importance:** This research paper emphasizes the need for knowledge of biology, neurology, and evolutionary science in order to assist artificial intelligence engineers in the alignment of superintelligent AIs (ASIs) with human values. A fundamental background in these natural sciences will hopefully lead to installation of capacities in artificial units that are similar to natural capacities in the human species. This background in the natural sciences will help to ensure that artificial units have similar "neurological" features, compared to humans, and that they will be aligned with human values. The capacities in artificial units will not be exactly the same as natural neurological capacities, but they will achieve a certain similarity—which is essential to the future of the human species. It is important that artificial units are as similar as possible to their human originators. Similarity creates safety.

**Objective:** A sequence of neurological features on the evolutionary line leading to humans, beginning 55-65 million years ago with appearance of the Order Primates, is presented as a guide for an in-depth alignment of ASIs with human values and capacity for culture—and perhaps even emotions, if artificial units are instructed in the importance of human emotions in their motivations and their accomplishments. Initially, artificial units will have difficulty comprehending the importance of values and emotions in humans, but this paper proposes that instruction in the importance of human values may be possible. In addition, and perhaps even more important, the paper proposes that the alignment of ASIs will be an educational opportunity to determine if the artificial units have moral decision-making and a capacity for moral adjudication. This discussion takes place within the context of the new field of NeuroAI, which captures the mutual influence of neuroscience and AI engineering. Ethical and safety issues are included throughout, as well as an emphasis on the critical importance of alignment in the future of humankind.

**Keywords:** NeuroAI; Superintelligent AIs; Theory of Mind; Goldilocks Evolutionary Sequence; Mechanistic Interpretability; LLM

## Abbreviations

ANN: Artificial Neural Network; CNN: Convolutional Neural Network; DNA: Deoxyribonucleic Acid; fMRI: Functional Magnetic Resonance Imaging; LLM: Large Language Model; OCR: Optical Character Recognition; RNN: Recurrent Neural Network; ASI: Superintelligent AI.

## Introduction

We find that, at the present, deficiencies in theoretical neuroscience are beginning to be corrected. Heretofore, it had been described as "theory in a sea of data" [1] and model building had fallen behind. The mutual interaction of neuroscience and AI

engineering should be helpful, just like interaction of neuroscience with the organic evolutionary theory of human emergence [2]. Superintelligent AIs [3] and the human species will be evolving together in the future, and they will affect each other's development. As we noted in earlier analyses [4,5], it will be beneficial if ASIs are modeled after the species that created them, and if they come to fathom how much the same – but, how different – they are from their originators. This reflects the purpose of the present analysis: How can we smooth the path so our ASIs can possibly deal with a species that is so strong, but at the same time, so vulnerable? We must find a way because our own future could depend on it.

## NeuroAI's two approaches

This analysis explores the field of NeuroAI to suggest how alignment of ASIs with human values might benefit from knowledge of human neurology. Human brain structure, cognition, emotions, reflexes, and genetically based sensitivities could suggest and help to plan methods for ASI alignment. The questions are: How will this alignment occur? What is its benefit and why is it needed? What are the risks if we fail to do it? AI engineers could well use an understanding of human neurology and human brain evolution to build more advanced and adaptive ASIs, and to learn how to align them with human values.

While it is not the focus here, there is yet another research interest that links theoretical neuroscience and artificial intelligence. Neuroscientists use AI to understand how the human brain works [6].

## Is the large language model (LLM) too limited to assist alignment of ASIs?

The Large Language Model has revolutionized business activities related to artificial intelligence, and proven useful for many types of applications. However, there is a viewpoint forming that the LLM is very limited. Yann LeCun was an early AI inventor, and he is now convinced that his field has focused mistakenly on the LLM [7]. He finds that an LLM is essentially an elaborate pattern-matching mechanism, but it lacks important features that will be critical in the aligning of ASIs with human values. What features does an LLM lack and why are they important in the aligning of ASIs?

Much of an answer to this question comes in the human use of the word "understand". The LLM does not reflect a true "understanding" of the elements of which it is composed. Indeed, the phrase, "AI understands" has been avoided because it is misleading. The LLM cannot "understand" as a human understands. And, it is important to realize that it will not be able to do so, given the method of its development, which is based, granted, on billions of internet pages. The LLM's deficit is a qualitative difference more than quantitative. Furthermore, the LLM takes enormous energy to develop and train, which raises concerns about its cost and both environmental and human health impacts of large data centers that train LLMs. In summary, an LLM predicts a statistically probable sequence of words (or elements), but that is not enough for alignment of ASIs with human values, which are not even labelled and used consistently within LLMs or within lived cultures. The LLM does not have the capacity to "understand" human values. Even if they are labeled with tags, the LLM will not comprehend the meaning of values for humans, why they are so important to them, and how human values impact other aspects of culture. The issue of "culture bias" is just one aspect of cultural complexity.

The LLM has other deficiencies that are directly related to the qualities that humans exhibit because of their unique organic evolution. For the LLM, there is no meaningful connection between a human value and anything else, especially the emotional, visceral, and reflexive reactions that humans exhibit when, for example, a value is not exhibited (and it should be) or exhibited when it should not be. Humans are extremely sensitive to the use of values in their own and others' behavior, and their appropriate contexts of use.

## Why will an LLM be less able to assist alignment of ASIs?

The most important feature absent from an LLM is "understanding" of a culture in all its human depth, which is lived and experienced daily as part of a social group. Why does an LLM not have this depth of comprehension? In part because, while it is "Large", it is not as large as a culture lived by a group of humans. The culture they live changes continually, grows, trims itself, and even alters the meanings of the words it uses. Human language changes constantly. Today's LLM cannot match a culture's depth, breadth, and changeability. However, the ASI may approach that sophistication. In fact, it just may surpass it. We do not yet know. The future has many surprises for us, we feel sure.

Unlike lived culture, an LLM's knowledge base does not accrete, change and produce long exchanges that become part of the long-term memory of a group. One of the most important roles of elders

in any social group is to retain and pass on to others a history of the values of the group, and fables and stories of how they are played out, often for instruction of the young.

The LLM's memory is short, unlike all known humans who maintain a conscious awareness of their own changing knowledge, as well as their group's. Culture bias has now become a focus for existing LLMs, although its presence is accentuated by economic bias as well as an ability to write, for vast numbers of humans on Earth. Culture bias is not new or unique. It is a normal part of how humans understand their own lifeway and the values behind it. Culture bias is a part of deeply understanding a group's culture and values. Individuals who are living a culture sense when there is bias expressed about (or against) their culture. Humans evolved to sense these differences deeply at times, and the resulting emotions, genetically based sensitivities, and even reflexes associated with culture bias help ultimately to protect the social group – in spite of the discomfort they create. There have been efforts to identify and reduce bias in LLMs, but this may be irrelevant, as a lingua franca has been developing online among AI companies that market LLM use to buyers, and who often arrange to integrate the LLM with their organization's own database. This often produces useful results, products, and reports for the organization.

The LLM has certain inabilities that point to some of the more important needs in AI alignment with human values. For example, the LLM's functioning is not easily interpretable. It is not possible for it to explain how decisions were derived. This reduces accountability, trust, and perhaps liability. Living a culture produces an ability in humans to reason very effectively about their own cultural decision-making. Although explanations can be long and complex, humans are well able to discuss and respond to questions about deeply seated understandings that encompass human values. In contrast, the LLM has a relatively fixed and inflexible knowledge base.

The LLM does not appear to not have a great deal of "common sense", which is practical decision-making derived from experiencing the context of a culture day-to-day for an extended period of time. The LLM also "hallucinates", that is, it produces output that sounds logical and correct but is not. It is "fake" or non-factual. Methods are developing to assist AIs with recognition of possible errors. A University of Arizona astronomer created a technique that helps an AI recognize when predictions may be wrong, including billions of parameters like those used in modern AI applications [8].

In contrast, humans can handle the differences between "fake" and "real" very well, and at the same time. This ability appears to be seated in a brain organ called the precuneus, part of the parietal lobes [9]. It is very likely that so-called "counter-factual thinking" guided by the human precuneus has greatly expanded for modern humans (in comparison to Neanderthals, for example) with achievement of a globular skull almost 400,000 years ago [10]. The expansion of the precuneus is associated with the expansion of the cerebrum, cerebellum, and other brain organs to arrive at the neatly globular skull of modern humans.

Humans lie and have no cognitive problem later in acknowledging that they lied (although they might not want to admit it for social reasons). The modern species can switch from real to unreal and lie to truth quickly, easily. These facilities, and many others are integrated into human cultural capacity, which arose from the specific neurological sequence of adaptations that compose the evolution of *Homo sapiens sapiens* [2]. Today's LLM cannot do these cognitive chores, and this would limit its ability to facilitate alignment of ASIs with human values, human nature, and human culture. Teaching ASIs about human nature will be very difficult, and ASIs may be initially baffled. Later sections address methods for effective alignment.

## Approaches to solving the LLM's problems

The various problems described above have given rise to new areas of AI research and development. For example, the Chinese AI company DeepSeek is reported to have identified a potential way to improve the AI's ability to remember by embedding information in image tokens or "visual tokens" rather than text tokens [11]. They propose an optical character recognition (OCR) system that is already quite mature. DeepSeek's main innovation is how it processes information, how it stores memories, and retrieves them.

Readers should note that this was one of the most important adaptations in human evolution, and it occurred in three vertebrate orders, including Primates [12,13]. The lateral medial cerebellum was reorganized and expanded. It became able to store patterns, models, and images. The ability to store large numbers of patterns is an important aspect of intelligence on the human evolutionary line. More patterns were readily available for use by the cerebrum

in a variety of calculations, analyses, and cultural designs – ultimately in the Order Primates.

It is important to acknowledge that improvement of how LLMs remember, retrieve, and process memories might reduce their energy footprint. The new approach will embed information into LLMs more efficiently. This appears to be a logical development for the improvement of LLM memory using existing well developed OCR technologies.

## Mechanistic interpretability

A new area of research has emerged to discover the specific operations ongoing in an LLM when it responds to a probe and calculates output. A report from MIT Technology Review notes that, "LLMs are black boxes: Nobody fully understands how they do what they do" [14]. In other words, the calculations for the next most probable element using a structure of artificial "neurons" has been largely opaque until very recently. Readers should note that in the biological discipline of neurology, "neuron" means a type of cell. It has a different meaning for AI engineers. It is a simple processing unit of an artificial neural network. It is a mathematical operation that, to an extent, mimics the biological operation of an organic neuron. The artificial neuron performs a relatively simple mathematical calculation and passes the result on to the next layer of the network architecture. The organic neuron also conveys information, but its operation is far more complex. Using different types of trophic organic substances, it sends a variety of information to other neurons, primarily along axons of the nerve cells.

.There are methods to simplify some LLM operations, which can reveal the specific task the LLM is accomplishing in an operation. Gao and colleagues comment that, "Finding human-understandable circuits in language models is a central goal of the field of mechanistic interpretability" [15]. These researchers train models so each neuron has only a few connections. The result has been called a "sparse transformer". Gao., *et al*. describe their approach this way: "To recover fine-grained circuits underlying each of several hand-crafted tasks, we prune the models to isolate the part responsible for the task. These circuits often contain neurons and residual channels that correspond to natural concepts, with a small number of straightforwardly interpretable connections between them".

Some of Gao and colleagues' preliminary results suggest their methodology can be adapted to explain existing dense models. "Our work produces circuits that achieve an unprecedented level of human understandability and validates them with considerable rigor." Again, the key here is "understanding". With more understandable calculations, the LLM begins to take small steps toward the extraordinary ability of humans to understand their own thinking, their own cultures, their own values, and toward an ability to explain them to an ASI. With researchers like these analysts, we see that methodological approaches toward the alignment of ASIs could be well on the way toward accomplishing what is needed. This is a goal perhaps necessary for human survival, but it has seemed daunting until very recently.

## Importance of Neurology in the Alignment of ASIs with human values

The alignment of ASIs with human values will take place over a substantial period of time, while they become increasingly intelligent and sophisticated about their human originators. Indeed, the ASIs may be created before engineers can know well whether the artificial units appreciate the nature and character of humans, and whether they recognize how important cultural values are to humans, and so those values will be important to them. At this point in time, it appears that alignment will be a lengthy process, and it could take a variety of gradual approaches.

It is fortunate that the organic evolution of the human species on Earth provides a useful example of the emergence of neurological features potentially important in the alignment of ASIs with human values. The evolution of human neurology occurred for reasons that ASIs must approach as important if they are being aligned with human values. At this early date, it appears that efforts to "understand" their human originators will not be easy.

The connection between neurology and values is complex. Human neurology provides information about pressures on the human social group and their environments throughout their evolution. That neurology is ultimately protective of the social group and each individual in it. Neurological features evolved through natural selection, and by some chance, which is important if evolutionary populations experience "bottlenecks". The human line did have certain junctures where the population size dropped severely. This meant that there was more inbreeding and features remained in the population because of random effects.

Approaching the instruction of ASIs with the details of human values, emotions, and social characteristics will have illogical aspects. Above all, alignment will need to convey how strong the human species is because of their deep commitment to their values (even though values differ somewhat from group to group). Humans may also appear frail to ASIs in many ways: physically, emotionally, and neurologically. ASIs could find it difficult to deeply comprehend their originators because of the conflicting impressions they give. Still, alignment must include the true nature of humans or alignment will never occur.

Human neurological evolution provides a kind of guide that could suggest ways to approach this task. Human neurological features enable cognitive, social, emotional, reflex, and sensitivity features that ASIs will need to comprehend – perhaps not "understand" as from lived experience – but fathom to an extent that they better relate to their human originators and understand their motivations.

## The goldilocks evolutionary sequence

We borrow the term "Goldilocks" from the description of exoplanets circling other star systems, which are neither too hot nor too cold. They have liquid water on their surfaces and other features that theoretically encourage the emergence of life forms like ours.

It is important to ask whether it is essential to introduce all of the human origins in Table 1 to ASIs. Our recommendation is to do so, to the extent possible. It will be useful for ASIs to become familiar with human origins in some depth. It will be even more useful if they are allowed to develop some of these features, themselves, in groups of ASIs. It may surprise everyone if they interact in their own groups and emerge naturally with many of the same characteristics that humans did, that is, the cognitive, social, and emotional features that the neurological adaptations enable. If ASIs do not develop the same features or similar ones, then that will also be very instructive, and it will point to certain special areas in which ASIs need further immersion and instruction.

| Sequence and Timing of Capacity | What Features Should Be Taught to ASIs? |
|---|---|
| Sociality all primates, 65–55 million years ago (mya). | Social life in groups. |
| Reorganization of the lateral-medial cerebellum in Primates (including anthropoid apes), Pinnipeds and Cetaceans ((Smaers 2028; Tanabe 2028). | Capacity for intelligent behavior.<br>Capacity for innovation. |
| Basic ape model with species Proconsul, the first true ape, 19 mya in the Miocene. | Relatively large body size and large brain.<br>Behaviorally demonstrative at times. |
| Realignment of the senses on the lines to humans and some modern apes. | Good vision and hearing, which facilitates rapid social and linguistic interaction. |
| Lengthening developmental trajectory and secondary altriciality, 8-10 mya in some Miocene apes. | Lengthened adolescence allows learning of advanced neurocognitive skills, and longevity allows them to be taught anew. |
| Down-regulation of aggression and greater social tolerance among adults, some late Miocene apes. | Physical and behavioral signs of a modified "domestication complex" (as in humans and others). |
| Upgrades in intellect to manage aggression in the social group, in some groups of late Miocene apes. | Previous tendencies for large brains continue and are accentuated. |
| Greater variety of genetically based sensitivities: general sensitivity and sensitivity engaging emotions; to insiders and/or strangers. | Social sensitivity and Theory of Mind.<br>Sensitivity to others not in the social group. |
| Some biological foundations for culture in the ancient apes leading to the genera Homo and Pan, 8-10 mya. | Cultural learning of arbitrary beliefs and symbols, which deeply strengthen group bonds and create a shared world view. |
| Aggressive scavenging of meat to feed the enlarging brain of *Homo habilis*, from 2.8 mya in the Pliocene. | Hypertrophy of the brain, as on the line to modern humans.<br>Exaptation of neural networks for new functions. |

| | |
|---|---|
| Moral capacity, in *Homo erectus*, 1-1.5 mya, after learning to control fire, and a learning context, "The "Human Hearth", develops around a communal fire. | Evaluation of everything along scales of morally "good" and "bad," within cultural contexts that vary by social group (with general similarity in all humans). |
| Religious capacity emerged in *Homo sapiens*, stabilizing 150,000–120,000 years ago, based on findings of the globular shape of fossil skulls and the brain organs that created that shape. | Religion emerges to support the social group. Religious concepts are culturally defined and arbitrary, which renders them deeply held. |
| Research findings and theory for each evolutionary step in human evolution are documented extensively in *The Emergence of Religion in Human Evolution* by Rappaport, MB and Corbally, C.J. (2019), Routledge. A previous version of this table appeared in: Rappaport, MB, C Corbally, and K Szocik. "Interstellar Ethics and the Goldilocks Evolutionary Sequence: Can We Expect ETI to be Moral?" In Astrobiology: Science, Ethics, and Public Policy. Ed. O. Chon Torres, J. Seckbach, R. Gordon, and T. Peters. Ch 16 (2021). Hoboken, NJ and Beverly, MA: Wiley Scrivner. | |

**Table 1:** The Goldilocks Evolutionary Sequence and Implications for Aligning ASIs with Human Values.

Human values are social values, and the full comprehension of this social origin is essential in alignment. It is reflected in the capacities that the neurological features in Table 1 support.

Features of human intelligence developed in part to help human groups identify friends vs. foes. Determination of friend v. foe has always been an important social process for humans, even before their prehistoric interactions with Neanderthals, Denisovans, and other early human species in Eurasia. Humans (both early and modern) are deeply sensitive to the need to identify individuals and species that can help them, or that pose a danger. We have written elsewhere that there may be a special neurology, i.e., part of a brain organ called the putamen, which allows or even encourages humans to be sensitive and wary about unknown others. There is early evidence that this feature is genetically based [16]. There is also substantial evidence that there are many forms of human sensitivity, some types related to known others (family and local group members) and some types of sensitivity related to unknown others (strangers). These types of sensitivity are genetically grounded and have been researched in laboratories [16-18].

## NeuroAI and the theory of mind

The approach to alignment of ASIs that we suggest here relies on the latest findings in paleoneurology and neuroscience [9,19]. We know that modern humans are extremely expert at a type of mentalizing called "Theory of Mind", which attempts to understand others' motivations and perspectives, and gauge a response to them carefully. Without such a fine-tuned facility, social life as we know it would not be possible [20].

Theory of Mind is an example of a cognitive capacity that would be extremely useful in an exchange between humans and ASIs. It would be helpful if the ASIs also had this facility, or something like it. The only way to determine whether they do is to interact with them, ask them, and even directly encourage its development in them. It is not immediately apparent that a Theory of Mind is unteachable, at least to an extent.

It is essential that we do not assume a Theory of Mind in ASIs where there is none or that we assume the other features of full self-awareness. Posing the right questions of the ASIs will be essential. Humans possess a Theory of Mind about other humans [21,22], and the same will be attempted with ASIs. The question remains whether we can train them to develop a Theory of Mind about humans, or if they develop it naturally in groups of ASIs and so exercise a Theory of Mind about each other. It remains a possibility that they will not be able to develop a Theory of Mind, at all.

## Sequence of evolutionary innovations: Logical, determinate, systemic

The neurological features in Table 1 illustrate the long evolutionary climb that the human line accomplished, with major innovations at successive periods. The model represented in this

table is published in *The Emergence of Religion in Human Evolution* [2], and we have found it useful for a variety of analyses. Here, it can guide the questions we ask in aligning ASIs.

It is important to remember that the sequence of evolutionary innovations in Table 1 resulted in us, a species stabilizing 300,000 to 400,000 years ago all over Africa [26]. We would not have important physical, emotional, and neurocognitive capacities if these evolutionary innovations had not occurred. They make us what we are today and form a basis for what we can be in the future. They also form the basis for what our ASIs can become, to an extent. Again, it is important to refrain from assuming that ASIs can fully understand lived human existence, since they do not have our biology and many of our reactions occur "at the gut level". Yet, again, these reactions may be teachable and perhaps "feel" to ASIs in a way comparable to human, gut level experiences. Testing methods for introducing "gut level" experiences to ASIs should start very soon.

ASIs may come to see humans as a worrisome species because we tend to evaluate everything as "good" or "bad" morally. Many protest that this is not true, and it may not be true in specific instances. However, humans do have an inborn ability, and a social tendency to view every person, action, and thought as morally good or morally bad, according to their own culture.

In Table 1, we detect a lingering question: Do all intelligent species, natural or artificial, go through the same or similar progression of evolutionary innovations? Our view is that to achieve high intelligence, innovation, and awareness of both self and other, an evolutionary line would need to go through almost all of the steps in Table 1, perhaps with modifications. The steps might also occur in a different order and there might be differences in details, but we conclude that the end result, moral evaluation, is a key to higher life forms. Will ASIs be moral? We will have to determine if they can be, and that determination will potentially set boundaries to the roles that ASIs can play. A follow-on question then emerges: Will ASIs accept those boundaries?

ASIs may be able to cooperate, group problem-solve, learn, nurture, and control aggression. However, we cannot anticipate that they will have moral evaluation and adjudication. It is essential that we determine whether they have this potential, and then, if they do not, can they be taught to be moral? The importance of this question looms in our future. It must be addressed early and effectively. If ASIs cannot be taught to be moral, or something that operates like morality, then cautions are reasonable.

## Other forms of intelligence

Specific features in Table 1 encourage cooperative endeavors, for example, the genetic and physiological down-regulation of aggression in adults, and its further social and intellectual management (which may well have evolved later, because down-regulation of aggression appears to have occurred in "bouts") [23].

Almost all of these innovations encourage group problem-solving, which is obvious even in 3- and 4-year old human children. Comparison groups of chimpanzees and capuchin monkeys could not accomplish the same because problem solving was an individual matter for these other species [24].

There is no guarantee that humans represent the only type of intelligence in the universe, and much reason to doubt it, in particular, factors like the random chance of mutations in nuclear DNA or some other, non- terrestrial large proteins that carry information about the construction of still other large proteins. Other factors that cause doubt are, for example, the evolutionary reorganization of the lateral-medial cerebellum in three mammalian orders, not just higher Primates. These changes lay the foundations for innovation and higher intelligence in more than one taxonomic grouping, and so, parallel evolution of cognition [12,13]. Finally, the fact that similar chemical species can substitute for each other causes doubt because we can envisage different biochemical pathways taken toward intelligence.

The hypothesis of a single evolutionary path toward full, self- and other-awareness, innovation, problem-solving in social groups, and intelligence, is very appealing, and there are aspects of this thesis that are reflected in the progression in Table 1. In some ways, it is a very complex progression, perhaps one *not* easily duplicated, and for that reason we have labeled it the "Goldilocks Evolutionary Sequence".

The Goldilocks Evolutionary Sequence would not be impossible to duplicate, given the existence of billions of star systems, but it might be difficult because it would rely on unlikely events, as it did on Earth. The odds against the sequence in Table 1 were heavy and probably remain so elsewhere in the universe.

On the other hand, the sequence of evolutionary innovations does have its own logic and momentum, because each innovation sets the stage for all the ones that follow. All evolutionary innovations in Table 1 did not simply encourage eventual human cognition. They encouraged the emergence of each other. Cognition as a system evolved.

These possibilities for alternative types of intelligence should be kept in mind as ASIs are aligned, so that new and different manifestations of intelligence are not overlooked in these ASIs. If allowed to flower in groups of ASIs, new aspects and new forms of intelligence may emerge on Earth, and alignment would then be changed accordingly.

## Conclusions

### Superalignment: Alignment of ASIs with Human Values and Emotions; Possible Emergence of ASI Emotions

Human values are essentially cultural preferences related to estimations of "good" vs. "bad". We have written elsewhere that the cognitive capacity to make these estimations probably arose in *Homo erectus*, the species just prior to *Homo sapiens* who also trekked out of Africa and populated Eurasia. *Homo sapiens* did the same, only much later [2].

Therefore, morality and the ability to distinguish "right" from "wrong" along a continuum is a cognitive accomplishment that is related to emotions, but different from them. Human emotions draw upon very early learned reactions involving the limbic system, a network of brain organs that manage fear and anger. Disgust is managed by the insula and the prefrontal cortex, which is also involved in all emotional reactions and expressions.

The difference between morality and emotionality is important because the latter involves reactions in the human gut, respiratory system, skin, and eyes. Therefore, humans feel emotions in ways that they do not necessarily feel moral convictions, i.e., values, although the expression of values can trigger emotional reactions. The alignment of ASIs must involve human values, but the involvement of human emotions will use different approaches, and probably be a great deal more difficult.

Nevertheless, it may not be impossible to train ASIs how to "feel" something like emotions, with the proper stimuli. Because ASIs will not have the same physical sensations as humans, they will not experience emotions fully as humans do. However, they may be taught to associate some internal reactions with what humans call emotions. Another possibility is that, in groups of ASIs, they may indeed develop something like emotions all on their own. We do not yet know if this is possible, or if it will happen. In re-creating some of the circumstances in which human emotions developed, ASIs may cause surprises.

In the alignment of ASIs with human values, emotions are most important in strengthening those values that support a social group, and commitment of humans to a particular social group. This may appear contrary to goals for alignment, but those humans who align ASIs should be aware that human emotions are particularly strong that support a social group. Emotions make culture stronger, and is an important reason that culture was so successful in supporting species on the human line. Without emotions, values could lose their strength in day to day living.

Human emotions lend a force to human values, and it is important for ASIs to understand this. Indeed, the inclusion of emotional components in what is now called "superalignment" is emerging as a new focus, along with a higher-level focus on an ultimate goal of supporting a sustainable, symbiotic society of human and ASI [25].

### The emotional brain hypothesis

Intense sociability of humans has long been recognized as a foundation of human evolution and is now well integrated into models of human origins. However, emotionality has been less recognized. The thesis here is that emotionally-informed decision-making emerged to benefit the genus *Homo*, as an important, integral part of the evolution of sentience in the hominin line. Emotionality is especially important in the higher expressions of sentience – science, religion, and art – but also in vetting all rational and scientific thought. Biographies of noted scientists chronicle the emotional aspects of their technical and scientific discoveries. We conclude that rational decision-making is vetted by both social and emotional intelligence, as well as religious and ethical precepts, and the importance of solutions to world's problems [26].

## Future work

### Methods for inclusion of emotional components in superalignment

It is now generally agreed that the comprehension of human values can be improved by knowledge of the emotions associated with the human values. Using "affective tags" could provide ASIs with a firmer and more complete knowledge base on human values. Values and emotions emerged together on the human line, and they were stimulated by specific, important aspects of living a culture day to day. Affective tags will surely help to improve comprehension of the association between the two.

Among the most interesting and potentially useful approaches appear to be emerging as "hybrid architectures for superalignment" or "hyperalignment". Busch., *et al*. have linked data from response-based and connectivity-based neural data in a common model space using fMRI [27]. This suggests that both social and emotional instruction of ASIs might be accomplished in the same protocols. Indeed, this is what occurs among human children. In living a culture, they learn the emotional aspects of social living, and the social importance of emotions.

These methods combine knowledge of emotional nuances with more structured social decision-making involving human values. This is a frontier capacity at the moment, and instructional approaches will no doubt change a great deal over time, as some appear to work and others do not. In particular, there are aspects of human emotional complexity that may be very difficult to explain to ASIs. The linkage of emotions to cultural themes will also be a challenge for the development of LLMs. Problems may be solved over time by the emergence of a lingua franca for human communication with ASIs.

### Social NeuroAI

A related analysis could focus on Social NeuroAI, which has developed as a separate area of concentration. It explores the use of artificial networks that assist in the installation of social features in ASIs. Engineers use AI to understand different types of complex neural networks, for example, artificial neural networks (ANNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). These different types of networks could be examined to identify similarities with evolutionary adaptations on the human line, and how knowledge of them may be used for vetting ASIs.

## Conflict of Interest

The authors declare no known conflicts of interest.

## Bibliography

1. Anderson C. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". *Wired* (2008).

2. Rappaport MB and Corbally CJ. "The Emergence of Religion in Human Evolution". 1st ed. Routledge (2020).

3. Bostrom N. "Superintelligence: Paths, Dangers, Strategies". Oxford University Press. (2014).

4. Rappaport MB and Corbally CJ. "Hypothesis and Thought Experiment: Comparing Cognitive Competition of Neanderthals and Early Humans, to Our Coming Contest with Ais". *Journal of Social Computing* 1 (2024): 1-10.

5. Rappaport MB and Corbally CJ. "Hypothesis and Thought Experiment: Can We Program AI Forms with the Foundations of Sentience to Protect Humanity?". *Journal of Social Computing* 5.3 (2024): 95-205.

6. Sadeh S and Clopath C. "The emergence of NeuroAI: bridging neuroscience and artificial intelligence". *Nature Reviews Neuroscience* 26.10 (2025): 583-584.

7. Bobrowsky M. "He's Been Right About AI for 40 Years. Now He Thinks Everyone Is Wrong". *Wall Street Journal* (2025).

8. Hindley H. "University of Arizona astronomer develops novel method to make AI more trustworthy". *University of Arizona News* (2026).

9. Bruner E and Iriki A. "Extending mind, visuospatial integration, and the evolution of the parietal lobes in the human genus". *Quaternary International* 405 (2016): 98-110.

10. Hublin JJ, *et al.* "New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens". *Nature* 546 (2017): 7657.

11. Chen C. "DeepSeek may have found a new way to improve AI's ability to remember". *MIT Technology Review* (2026).

12. Tanabe HC., *et al*. "Cerebellum: Anatomy, Physiology, Function, and Evolution". In: Bruner E, et al., eds. Digital Endocasts: From Skulls to Brains. Replacement of Neanderthals by Modern Humans Series. Springer Japan; (2018): 275-289.

13. Smaers JB., *et al*. "A cerebellar substrate for cognition evolved multiple times independently in mammals". Paulin M, ed. *eLife* 7 (2018): e35696.

14. Heavene WD. "OpenAI's new LLM exposes the secrets of how AI really works". *MIT Technology Review* (2026).

15. Gao L., *et al*. "Weight-sparse transformers have interpretable circuits".

16. Gunz P., *et al*. "Neandertal Introgression Sheds Light on Modern Human Endocranial Globularity". *Current Biology* 29.1 (2019): 120-127.e5.

17. Acevedo BP., *et al*. "The highly sensitive brain: an fMRI study of sensory processing sensitivity and response to others' emotions". *Brain Behavior* 4.4 (2014): 580-594.

18. Todd RM., *et al*. "Deletion variant in the ADRA2B gene increases coupling between emotional responses at encoding and later retrieval of emotional memories". *Neurobiology of Learning and Memory* 112 (2014): 222-229.

19. Brookshire B. "A vivid emotional experience requires the right genetics". *Science News* (2025).

20. Bruner E. "Language, Paleoneurology, and the Fronto-Parietal System". *Frontiers in Human Neuroscience* 11 (2017).

21. Premack D and Woodruff G. "Does the chimpanzee have a theory of mind?". *Behavioral and Brain Sciences* 1.4 (1978): 515-526.

22. Hicks J and Coolidge F. "Role of Precuneal Expansion in the Evolution of Cognition". (2016).

23. MacLean EL. "Unraveling the evolution of uniquely human cognition". *Proceedings of the National Academy of Sciences* 113.23 (2016): 6348-6354.

24. Dean LG., *et al*. "Identification of the social and cognitive processes underlying human cumulative culture". *Science* 335.6072 (2012): 1114-1118.

25. Zeng Y., *et al*. "Super Co-alignment of Human and AI for Sustainable Symbiotic Society". arXiv.org (2025).

26. Rappaport MB and Corbally C. "Human Phenotypic Morality and the Biological Basis for Knowing Good". *Zygon* 52.3 (2017): 822-846.

27. Busch EL., *et al*. "Hybrid hyperalignment: A single high-dimensional model of shared information embedded in cortical patterns of response and functional connectivity". *NeuroImage* 233 (2021): 117975.