

phhh

by Harith Albayati

Submission date: 26-Nov-2020 03:34PM (UTC-0700)

Submission ID: 1345286980

File name: Philosophy_of_Artificial_Intelligence_2.docx (36.63K)

Word count: 3558

Character count: 20789

Philosophy of Artificial Intelligence

³
Khalid Alattas, Department of Computer Science and Artificial Intelligence, College of
Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia.
kaalattas@uj.edu.sa.

Abstract

¹
The defense sphere of artificial general intelligence (AGI) is growing exponentially. However, there is an under the definition of the character of human beliefs with which future AGI should be associated. Different AGI protection scholars have formulated numerous hypotheses regarding the existence of human beliefs, but contradictions exist. This text provides an analysis of what AGI protection scholars, up to the beginning of 2019, have written about the essence of human beliefs. It is generally advised to use a theory classification system, where the ideas are evaluated following the degree of their sophistication and size of behaviorists-internalists, equally because of the scope of their generality-humanity. We propose that several well-supported hypotheses indicate that it is difficult to describe the character of human beliefs, and a few meta-level theories are needed.

Keywords: Artificial Intelligence, human values, Artificial General Intelligence.

1. Introduction

¹
AGI alignment, i.e. the assumption that future advanced AGI will acquire human values, is the most current approach to artificial general intelligence (AGI) defense, and therefore its utility role will be compatible with human values. This partnership consists of three elements: addressing AGI with the target system, the value learning process, and "human values." All three are mutually related as various concepts assume different ways of presenting and learning about the essence of human values. AGI alignment analysis is usually described in the context of mathematical methods aimed at offering accurate ways to learn human values (e.g. inverse reinforcement learning). It is clear to most AGI protection researchers that the idea of "individual principles" is unclear, and before it can be included in (mostly mathematical) AGI alignment models, this concept should be more formalized. In other words, "AGI-conscious" ideas of human value are those that were expressly developed to assist research into AGI alignment. Most of the latest psychological hypotheses of human

values are textual, informal, and underdefined, so before they can be extended to AGI defense, they need some adaptation.

In certain situations, the philosophy of human values cannot be differentiated from how a potential AGI is supposed to extract values, such as approval-directed AGIs. Multiple hypotheses have also been proposed by some researchers (notably, Armstrong); including a connection to the researcher is not meant as a suggestion that the researcher adheres to precisely this theory.

Principles of the grouping of theories

- 1) Explanation of the complexities of values (not the complexity of the theory): Certain theories of human values suggest that human values are very simple, e.g. that there are only two drives: survival and reproduction; or that there is only one ability to optimize pleasure; or that there are only nine essential feelings. Other ideas show that human beliefs, such as a continuum of interaction between all meanings and incentives, are very complex.
- 2) The level of "behaviorism": most ideas are drawn to one of two poles: "internalist" theories that believe that values exist, but are suppressed within the human brain, and "behaviorist" theories that assume that values only manifest in human behavior (such as acceptance or choice). The principles and the process of their extraction (e.g. approval-directed AGI) are combined by behaviorist theories of human values (generally). Internist ideologies, on the other hand, usually hold that beliefs exist independently from the way they are studied.

According to the level of abstractness, another way to describe hypotheses regarding human beliefs is: certain theories may be generalized to any conceivable mind and thus do not take input from human psychology or neurophysiology. These theories are computationally abstract and can also contain implicit beliefs concerning some of the human mind's properties: equilibrium, harmony, continuity, etc. Human-centered theories depend on established human-minded theories. So there is research question in this paper that What is the future in development of emotional intelligence in the field of artificial intelligence?

2.Literature Review

Yudkowsky is a pioneer in the field of AGI defense, and he developed the concept of the "complexity of values" among many other things; basically, that any brief verbal explanation does not comprehend the complexity of the effects we seek. In 'Complex Value Structures are Needed to Realize Valuable Futures,' he outlined his criticism of simplistic wishes as to the proper portrayal of desirable results. He also introduced the idea of "fragility

of principles" in the same post, e.g., if one digit is incorrect in a phone number, the call will go to a different user. The principle of Coherent Extrapolated Volition (CEV) is another important contribution from Yudkowsky. "In his" Complex Principles "essay, he wrote:" We should strive to describe normativity without unnecessary lack of will (failure of self-control) not by our immediate present impulses but by our reflective equilibrium, what we will want in the limit of perfect knowledge, the freedom to weigh all alternatives and claims, and perfect self-knowledge.

The Value (Arbital, 2015) begins with a definition: "In the sense of value alignment as a subject, the term 'value' is a speaker-dependent variable that indicates our ultimate objective - the property or meta-property that the speaker wants or should want to see in the outcome of intelligent life originating on Earth. Most psychologists, and individuals in general, interpret "human ideals" differently, since, in their day-to-day interests, most people do not think for the distant future of mankind.

A list of potential points of view on the existence of human values is also provided in the Value article on values and can be summarized as;

- Reflective balance. Provided "more factual knowledge, time to consider more knowledge, better self-awareness, and better self-control," what one should want.
- Regular wishes. "A view at the object level that defines meaning with characteristics that we currently find rather attractive, enjoyable, fun, and preferable.
- Instant items. "For example, 'Cure cancer'"
- Principle of deflationary moral mistake. This mostly concludes in reality with an "immediate goods" philosophy, plus some views related to the debate on value selection
- Clear goal. "Price, for any X, can easily be associated with X."

In Defining Human Values for Value Learners, Sotala (2016) analyzes human values. He listed several problems with the basic model of human values as a utility function in this article: the value function utility model has trouble dealing with internal contradictions and higher-order preferences, the utility function model of valuation lacks the internal perception of the person, the value model of the utility function does not model shifting values, and the value utility function model does not have a way to generalize to new ones from our current values.

He proposes the following description to address this problem: "... human principles are concepts that abstract over circumstances in which we have earlier earned benefits, making those concepts and the situations associated with them respected for their own sake." A further implication is that the value function could be at least partly embedded in the

influence of the different concepts, as people seem to instinctively find different mental concepts to be correlated with the result (the subjective perception of a sensation or emotion, perceived as either positive or negative).

Sotala suggests helpful parameters at the end of the essay for estimating the authenticity of some theory of human values. He states why a hypothesis like this should be: realistic psychology, compatible with separate variants, testable, integrated with proven hypotheses, suited for exhaustively modeling different values, suitable for modeling interpersonal contradictions and impulses in greater order Suitable for modeling principles shifting and evolving and should finally suitable for generalizing from current to new values

In their article Mammalian value systems, "Sarma and Hay (2019) note that" [a] agent using Inverse Reinforcement Learning or Bayesian Inverse Planning to learn and refine its human values model by observing our actions must begin with some very rough or preliminary initial assumptions about the existence of the values it is attempting to learn. The definition of the basic mammalian value structure is based on the work of Panksepp and Biven, who "categorize the above informal list into seven motivational and emotional mechanisms common to mammals: searching, anger, terror, lust, caring, panic/grief, and play." Sarma and Hay then add the word "neural value correlate," which is some subcortical, according to them. That essay concludes: they argue that what we refer to as human beliefs in colloquial terminology can be informally decomposed into 1) mammalian values, 2) human thought, and 3) human social and cultural evolution over many centuries."

Sarma and Hay (2018) find out in AI Protection and Reproducibility: Building Robust Foundations for Human Values Neuropsychology,' that the reproducibility problem in psychology renders it challenging to define the right theory of human values, so immediate intervention to that end is required to ensure the safety of AGI. "Integrative Biological Modeling, Neuropsychology, and AI Defense," is the newest report by Sarma, Safron, and Hay (2019), in which they propose developing improved biological models of animal minds to gain a better understanding of the essence of motivation.

Especially enticing is the possibility of a 'prior' for studying human principles, which could be reasonably easily taught by a non-human source. This is because Bayesian approaches are focused on multiple learning techniques, which require some previous models to start working. This would make the learning process better because any animal is likely to oppose dissection of the brain or other non-ethical merit learning practices (except maybe after natural death). However, since there are undoubtedly other models of animal

motivation, the selection of seven primary emotional-motivational mammalian traits may appear random. Why are mammals, too, but not primates or vertebrates?

¹ In an article by Muehlhauser et al (2012), "The Singularity and Computer Ethics," they show ¹ that if built in a strong optimizer, certain (and potentially all) established moral philosophies are dangerous. The state and address the theme in Section 5.1 of that article that "[h] people do not know their beliefs," ¹ based on an experiment in which participants ⁵ explained their affinity for faces not selected by the researcher. They note that "cognitive science shows instead that our perception of our preferences is much like our understanding of the desires of others: assumed and sometimes inaccurate."

¹ Sezener indicated that human values are an arbitrary dynamic incentive mechanism in the article ¹ Inferring human values for stable AGI architecture (Can & Sezener, 2015). The core principle of Sezener is the use of Solomonoff induction to find the simplest combination of two programs, one of which encodes the compensation function of an agent and the other, based on a measurable sequence of acts and observations, encodes the agent itself. This is analogous to the approach of Armstrong to present humans as a (p, R) preparation ¹ algorithm and reward pair, and then to find the simplest such pair that describes measurable actions using complexity considerations.

In inverse reinforcement learning (IRL), Sezener also wrote of implicit assumptions: Significant opinions of experts recommend using approaches ¹ similar to IRL for the learning of human beliefs. The existing IRL techniques, however, are limited and, because of their long list of hypotheses, should not be used to assume human beliefs. For eg, the world is generally presumed to be stationary, completely measurable, and often understood in most IRL methods; the agent's strategy ¹ is assumed to be stationary and optimal or near-optimal; the compensation function is often ⁷ assumed to be stationary, and the property of Markov is assumed. For restricted motor control activities such as gripping and manipulation, such conclusions are rational, but if we aim ¹ to learn high-level human values, they become impractical.

We interpret Sezener's approach's key problems as his statements that:

1. The right description of a human incentive mechanism (what about implicit or parasitic actions?) is behavior and only behavior.
2. Function Compensation = values;
3. Ignores internal inconsistencies in the model;
- ¹ 4. The model is incomputable since it is based on the incomputable AIXII of Hutter.

5. The flexibility of both the incentive function and the agent is believed to have a "free lunch."

In his article Friendly AI by Ontology Auto-generation (2018), Maxwell wrote: "If an AI is to be Friendly, it must function based on an ontology that is capable of communicating our values," and "[r]egardless of the ontology auto-generation algorithm chosen, it is almost inevitable that the original auto-generation would either (a) catch human values with inadequate fidelity or (b) contain too many." In other terms, without ontology, Maxwell proved that meanings cannot exist and are closely related to them. And if AGI is used to produce ontology, the process of seeking values may not be straightforward.

In "Avoiding Unintended AI Actions," Hibbard (2012) wrote that an officer "should ask model human d to convey a utility value between 0 and 1" for the policy to measure policy. This may be considered a "human model counterfactual acceptance," which considers all potential consequences of behavior. In the case of Hibbard, secure AGI consists of two levels: the first generates a world model (which involves all individuals and their ways of interacting or reacting), and the second measures how people will respond to hypothetical future histories in the model. Sezener is dismissive of Hibbard: a failure of this technique, though, is that what human models think they value and what they value can also be distinct.

Yampolskiy is also dubious that it is possible to formalize human values: "human values are complex and contradictory and can never be understood/programmed into a computer." Suggestions to solve this challenge enable civilization to be transformed into something it is not, and therefore to break it by definition as per the thoughts by Roman Yampolskiy on AI Protection Engineering (Muehlhauser, 2013). He suggests a solution in the essay "Personal Universes: A Solution to the Multi-Agent Value Alignment Dilemma (Yampolskiy, 2018) to avoid the complicated problem of aggregating value in a personal universe that will "optimally and dynamically change to align their values and desires [humans]." In other words, potential superintelligent AI would create a most fascinating simulation of a personal universe. This would eliminate the need for different individuals to aggregate different values, but it also involves choosing the most appropriate values within an individual.

In "Formally Stating the AI Alignment Dilemma," Worley (2018) wrote about the need for artificial general intelligence (AGI) alignment to take into consideration the mental phenomenon of consciousness. He also addresses his belief that the instruments of phenomenology (a field of philosophy that examines internal objects within consciousness) should be used to explain human beliefs."He clarifies his stance in a private communication:"

... my opinion is that ideals are inextricably connected to the nature of consciousness and they derive from our self-aware awareness. This implies that I think values have a basic, fundamental structure and also that values within that simple structure are rich with depth in their substance. This view also inherently implies that values from behavioral approaches are not entirely discoverable and that there is still a secret, internal component that might not even be accessible by the agent itself.

A general safety-oriented AI development model "was written by Dai (2018) where the production of AGI is an interactive process inside a human-AI team:" "Start with a team of one or two individuals with access to zero or more AIs (initially as assistants) (researchers, programmers, coaches, and/or overseers). The human / AI team creates a new AI in each round and introduces it to the team, and repeats this until AI technology sophistication is reached. Safety/alignment is maintained by the team providing a set of safety/alignment properties that the production process retains inductively. "Some AI protection concerns are likely to have counterparts in humans," Dai also wrote about the potential vulnerability of human values. AI designers and security scientists should not begin by pretending that humans are safe.

3. Methodology

The approach of the study was qualitative and as a result data was primarily dependent on literature by specialists in the AI spectrum. The analyzed literature was sourced from peer reviewed articles on technology and the ethics and future of AI. A segment of the data were presented in graphs to effectively explain the presented narratives.

4. Results

Understanding the philosophy of AI and the relation to the emotive quotient of human beings is a qualitative analysis of different literature by experts in the field. In the current literature, there are many internally coherent explanations of human beliefs proposed by AGI protection scholars. The ideas, though, are not quite consistent with each other, considering their internal continuity, and offer a wide variety of opinions.

There are also other "possible hypotheses," that is, methods that have not yet been put forward by any researcher (to the best of our knowledge), but that could be generated based on the same principles as other theories. One is to believe that all human principles are results of evolutionary fitness and can be extracted in the same fashion as Omohundro's AGI fundamental drives from basic evolutionary considerations (Omohundro, n.d.). This could describe the most fundamental human drives such as survival, gender, status-seeking, and instinct of discovery.

Another such idea is that, first, an AI Oracle can read the current psychological literature, pick the best theory of mind, and establish its framework based on that theory of human values (Armstrong, Sandberg, & Bostrom, 2012). In a corresponding article, we will discuss such a hypothesis. We may infer from this that internal consistency, as well as scientific support and comprehensive literature, is not adequate to provide us with a "true" theory of human values, since such support may also be given to other alternative theories. There is a need for a form of meta-theory of human beliefs, often related to their learning methods.

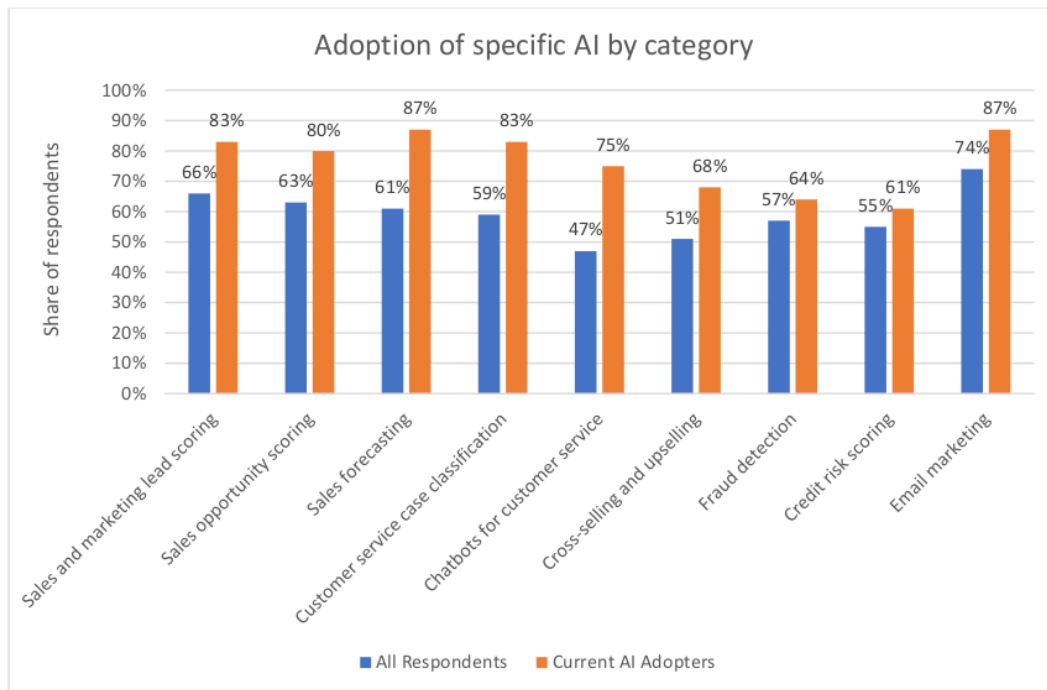


Figure 1: How AI has been adopted by various sectors

Current AI Adopters	Share of Respondents (%)	
Sales and Marketing Lead Scoring	66	83
Sales opportunity scoring	63	80
Sales for forecasting	61	87
Customer service case classification	59	83

Chatbots for customers services	⁴ 47	75
Cross-selling and up-selling	51	68
Fraud Detection	57	64
Credit risk scoring	55	61
Email marketing	74	87

Table 1: shows the categories for AI adoption

Figure 1 shows how AI has been adopted in various sectors in the economy. Evidently, the sectors highlighted are some that require the use of human intelligence for effective operations thus the need and question for the integration of emotive intelligence in AI.

5. Recommendations and Future Direction

There is a need for the utilization of AG in the generalization of human cognitive abilities. In this regard, when the AI systems are faced with challenging tasks, the AGI framework will be more oriented to tackling without much difficulty. Examples of such systems will include self-driving cars with expert super computers. Making AGIs as smart as humans will requires the customization of AI-based bots to enable real-time monitoring for processes such as marketing; thus, the AGU frameworks will be able to replicate overly similar intelligence. Future direction should also underlie the usage of AGI systems in the personification of human tasks. In this regard, the smart AI-powered bots can imitate human thoughts and dreams, especially for the future. As a result, the AGI systems can be used to comprehend human-robot interactions.

6. Conclusion

We have seen in this report that various AGI protection scholars have proposed different, often conflicting, hypotheses regarding the existence of human values. A system of hypothesis classification was proposed, ¹where the theories are tested according to the degree of their complexity and scale of behaviorists-internalists, as well as the level of their generality-humanity. We propose that some well-supported hypotheses indicate that ⁶it is difficult to describe the essence of human values, and some meta-level hypothesis is needed.

References

- Arbital. (2015). Value. Retrieved November 4, 2020, from arbital.com website:
https://arbital.com/p/value_alignment_value/
- Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines*, 22(4), 299–324:
<https://doi.org/10.1007/s11023-012-9282-2>
- Can, E., & Sezener. (2015). *Inferring human values for safe AGI design*. Retrieved from
http://agi-conf.org/2015/wp-content/uploads/2015/07/agi15_sezener.pdf
- Dai, W. (2018). A general model of safety-oriented AI development - LessWrong. Retrieved November 4, 2020, from www.lesswrong.com website:
<https://www.lesswrong.com/posts/idb5Ppp9zghcichJ5/a-general-model-of-safety-oriented-ai-development>
- Hibbard, B. (2012). Avoiding Unintended AI Behaviors. *Artificial General Intelligence*, 107–116. https://doi.org/10.1007/978-3-642-35506-6_12
- Maxwell, J. (2018, March 7). Friendly AI through Ontology Autogeneration. Retrieved November 4, 2020, from Medium website: <https://medium.com/@pwgen/friendly-ai-through-ontology-autogeneration-5d375bf85922>
- Muehlhauser, L. (2013, July 16). Roman Yampolskiy on AI Safety Engineering. Retrieved November 4, 2020, from Machine Intelligence Research Institute website:
<https://intelligence.org/2013/07/15/roman-interview/?fbclid=IwAR2AOyUc0JEySlgclbwoNcYzL7RzeUnyOFerHqQEgIEWGfuZkGCZQYCIWOg>
- Muehlhauser, L., Helm, L., Eden, A., Søraaker, J., Moor, J. H., Steinhart, E., ... Springer. (2012). *Intelligence Explosion and Machine Ethics*. Retrieved from
<https://intelligence.org/files/IE-ME.pdf>
- Omohundro, S. (n.d.). *The Basic AI Drives*. Retrieved from
https://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf
- Sarma, G. P., & Hay, N. J. (2019). Mammalian Value Systems. *ArXiv:1607.08289 [Cs]*. Retrieved from <https://arxiv.org/abs/1607.08289>
- Sarma, G. P., Hay, N. J., & Safron, A. (2018). AI Safety and Reproducibility: Establishing Robust Foundations for the Neuropsychology of Human Values. *Developments in Language Theory*, 507–512. https://doi.org/10.1007/978-3-319-99229-7_45
- Sarma, G. P., Safron, A., & Hay, N. J. (2019). Integrative Biological Simulation, Neuropsychology, and AI Safety. *ArXiv:1811.03493 [Cs, q-Bio]*. Retrieved from <https://arxiv.org/abs/1811.03493>
- Sotala, K. (2016). *Defining Human Values for Value Learners*. Retrieved from
<https://intelligence.org/files/DefiningValuesForValueLearners.pdf>
- Statista. (2020). Global AI software market growth 2019-2025 | Statistic. Retrieved from Statista website: <https://www.statista.com/statistics/607960/worldwide-artificial-intelligence-market-growth/>
- Worley, G. (2018). Formally Stating the AI Alignment Problem. Retrieved from medium website: <https://mapandterritory.org/formally-stating-the-ai-alignment-problem-fe7a6e3e5991>
- Yampolskiy, R. (2018). *Personal Universes: A Solution to the Multi-Agent Value Alignment Problem*. Retrieved from <https://arxiv.org/ftp/arxiv/papers/1901/1901.01851.pdf>

ORIGINALITY REPORT

35%

SIMILARITY INDEX

35%

INTERNET SOURCES

3%

PUBLICATIONS

2%

STUDENT PAPERS

PRIMARY SOURCES

1

philpapers.org

Internet Source

33%

2

Submitted to Deakin University

Student Paper

1%

3

Muhammad U. Ilyas, Muhammad Murtaza Khan,
Sajid Saleem, Jalal S. Alowibdi. "Chapter 45
IoT-IaaS: A New Public Cloud Service to
Unleash Innovation in the IoT Space", Springer
Science and Business Media LLC, 2019

Publication

1%

4

www.salesforce.com

Internet Source

1%

5

"Singularity Hypotheses", Springer Nature, 2012

Publication

<1%

6

iopscience.iop.org

Internet Source

<1%

7

Lecture Notes in Computer Science, 2015.

Publication

<1%

Exclude quotes Off

Exclude matches Off

Exclude bibliography On