Journal of Computer Science

Review

An Overview of Artificial General Intelligence: Recent Developments and Future Challenges

[1]Khalid Alattas, [2]Ahmed Alkaabi and [3]Alanoud Bandar Alsaud

[1]Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia
[2]Electrical and Computer School, RMIT University, Australia
[3]Administration College, British Institute of Economics and Political Science, UK

Article history
Received: 01-12-2020
Revised: 10-02-2021
Accepted: 09-03-2021

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

Abstract: The defense sphere of Artificial General Intelligence (AGI) is developing exponentially. Notwithstanding, there is an under the definition of the character of human beliefs pertaining to AGI associations. Distinctive AGI protection scholars formulated numerous hypotheses regarding the existence of human beliefs, but contradictions exist. This study 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 consensus mankind. We propose several well-supported hypotheses to indicate the difficulty of describing the character of human beliefs and a few meta-level theories are needed.

Keywords: Artificial Intelligence, Human Values, Artificial General Intelligence

Introduction

AGI sequence, that is, the assumption raising subsequent advancement AGI will acquire human qualities, is the most current approach to Artificial General Intelligence (AGI) defense. Therefore, its usefulness is compatible with human qualities. This partnership embodies three aspects: addressing AGI with the target framework, the esteem learning process and "human values." Every one of the three is commonly connected to various concepts assume discrete ways of presenting and learning about human values' essence. 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 apparent to most AGI protection researchers that the concept of "individual principles" is unclear and before it can be included in (mostly mathematical) AGI alignment models, this concept should be more formalized. By way of expression, "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 qualities are textual, casual andunder defined, so before they can be extended to AGI defense, they need some adaptation.

In certain situations, human values' philosophy 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 to suggest that the analyst sticks to precisely this hypothesis.

Principles of the Grouping of Theories

1) Explanation of the complexities of values (not the multifaceted nature of the hypothesis): Certain speculations of human qualities suggest that human qualities are exceptionally straightforward, for example, that the urges are two: Endurance and generation; or that there is just one ability to optimize pleasure; or that there are only nine essential feelings (Almeida *et al*., 2020). Other ideas show that human beliefs, such as a continuum of interaction between all meanings and incentives, are complex
2) The degree of "behaviorism": Many ideas are drawn from either of the poles: "Supporter of internalism" thesis, which believes that virtues are

present and are suppressed within mankind wit, together with "behaviorist" thesis that supposes that values are manifested in an individual's behavior. The principles and processes concerning their beliefs (e.g., approval-directed AGI) are combined by behaviorist theories of human values (generally). On the other hand, internist ideologies usually hold that beliefs exist independently from the way they are studied

According to the abstractness level, another way to describe hypotheses regarding human beliefs is certain theories may be generalized to a conceivable mind and in this manner do not take contribution from human brain research or neurophysiology (Almeida *et al*., 2020). These hypotheses 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 a research question in this study: What is the future in the development of emotional intelligence in artificial intelligence?

Contribution of the Study

In this study, we review the historical development of artificial general intelligence, artificial intelligence, machine learning and data science. We compare these developments with human intelligence and abilities. Also, the research looks at future possibilities in Artificial General Intelligence (AGI), while analyzing ethical dilemmas posed by future evolutions. The focus of the study is to analyze the evolution processes historically and in the future for a lay man's better understanding of artificial general intelligence and data science.

Gap in Literature

There has been much literature analysis the artificial general intelligence advancements. However, very little has been covered to analyze Deep Learning (DL) as technological evolution. It is essential to note that the advancements of DL is a vast topics in and of themselves, each carrying with them a many relevant subjects.

Application of Artificial General Intelligence

AG is applied in many instances; it has been adapted to complement our everyday lives. In this case, we look at Machine Learning (ML). Machine Learning techniques and data sets can be classified in two groups, that is, linear and nonlinear (Vita-More, 2020). A linear data pattern is not complicated which can be further grouped using a linear function to perform classification. Multiple algorithms have been established to fit many linear regressions, logistic regression, classification and regression trends and anchor vector machine. Nonlinear functions algorithms cannot be grouped using linear methods. Similar to other methods of analysing data and management, nonlinear data associations may offer more challenges to ML.

Although multiple algorithms have been established to perform linear data, the nonlinear nature of most data sets still poses a challenge for ML. For instance, the decision trees, k-nearest neighbors and anchor vector machine. This research paper entails a literature review, methodology and results, recommendations for future research and conclusion.

Literature Review

Yudkowsky is among the pioneers in AGI defense and he developed the concept of the "complexity of values" among many other things; basically, any brief verbal explanation does not comprehend the complexity of the effects we seek. In 'Web Virtue Structures Needed to Attain Important Futures,' he outlined his criticism of simplistic wishes regarding the proper portrayal of desirable results. He also introduced the idea of "fragility of principles" in the same post; for instance, if any of the digits is incorrect in telephone dial, the conversation proceeds with an alternative user. The principle of Coherent Extrapolated Volition (CEV) is another important contribution from (Yudkowsky, 2011). "In his" Complex Principles "essay, he wrote:" We should strive to describe normativity without unnecessary lack of will (failure of self-control), not through the available present impulses but using one's reflective equilibrium, we will require to curtail good know-how, freedom of weighing available alternatives together with claims and good know-how.

In an article by (Muehlhauser, 2013), "The Singularity and Computer Ethics," they show 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," based on an experiment in which participants 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."

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 behavior consequences. In Hibbard's case, 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 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, 2019) 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 super-intelligent AI would create a fascinating simulation of a personal universe. This would eliminate the need for different individuals to different aggregate values, but it also involves choosing the most appropriate values within an individual.

The Value (Arbital, 2015) begins by defining: "According to the sense of principle existence as the main idea, term 'value' is dependent with the speaker as a variable which points to one's main objective - the assets or meta-assets which a person might require to see in the result of intelligent life coming on Earth. Many individuals interpret "human ideals" differently since, in their day-to-day interests, many individuals do not think for the distant future of humankind.

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

- Reflective balance. Provided there is adequate facts about know-how, period to look into additional know-how, advanced self-knowledge and advanced control of oneself, what one should want
- Regular wishes. "A view at the object level defines meaning with characteristics that we presently find rather attractive, pleasing, 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"

Sezener indicated that human values are an arbitrary dynamic incentive mechanism in the article inferring human values for stable AGI architecture (Sezener, 2015). The core principle of Sezener is the use of Solomon off 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 Armstrong's approach to present humans as a (p, R) preparation 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 similar to IRL to learn human beliefs. However, the existing IRL techniques are limited and, because of their long list of hypotheses, should not be used to assume human beliefs. E.g., the world is generally presumed to be stationary, completely measurable and often understood in most IRL methods; the agent's strategy is assumed to be stationary and optimal or near-optimal; the compensation function is often 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 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
4. The model is incomputable since it is based on the incomputable AIXII of Hutter

In Defining Human Values for Value Learners, (Sotala, 2016) analyzes human values. He recorded a number of problems with the basic model of human values as the functional usefulness in this article: The value function utility model has trouble coping with inner contradictions and increased preferences of orders, the model of the utility function of valuation lacks the inner perception of a person, the value structure of the utility function does not structure shifting principles and the principle usage structure does not have a method of generalizing to new ones from one's current principles.

He proposes the following description to address this problem: "... Human principles are concepts that abstract over circumstances in which we have earlier earned benefits, causing the techniques together with the consequences concerning them respected for their benefit." Moreimplications show the value function could be partly embedded in the different concepts' influence. People seem to instinctively find different mental concepts correlated with the result (the subjective perception of a sensation or feeling, looked at as either positivity or negated).

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 ought to be: Real psychology, flexible to separate variants, able to be tested, transformed with proven hypotheses, fit for exhausting modeling alternating principles, 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.

Sarma *et al*. (2018a) find out according to AI Protection and Duties: Building Strength Lay grounds 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 *et al*., 2018b), in which they propose developing improved scientific structures of mammal thinking to gain advanced knowledge of the essence psyche.

Especially enticing is the possibility of an 'advance studying of mankind principles that can be reasonably simply taught by a non-humankind origin. This is because Bayesian approaches are focused on multiple learning techniques, which require some previous structures to start working. This would cause learning stages better because any mammal is likely to reject the mind's dissection or other non-morally merit educative practices. However, since there are undoubtedly other animal motivation models, the selection of seven primary emotional-motivational mammalian traits may appear random. Why are mammals, too, but not primates or vertebrates?

In his article Friendly AI by Ontology Auto-generation, Maxwell (2017) wrote: "If an AI is to be Friendly, it must function based on an ontology that is capable of communicating our values," and "[r]regardless 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. Furthermore, if AGI is used to produce ontology, the process of seeking values may not be straightforward.

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 values have a basic, fundamental structure and 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 moreInitially as Assistants (AIs) (researchers, programmers, coaches and overseers). The human/AI team creates a new AI in each round, introduces it to the team and repeats this until AI technology sophistication is reached. The team maintains safety/alignment, 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 (2018), also wrote about the potential vulnerability of human values. AI designers and security scientists should not begin by pretending that humans are safe.

In their article Mammalian value systems, "Sarma and Hay (2017) note that" [a] agent utilizing Indirect Reinforcement education or Bayesian Indirect Planning to gain knowledge and better its mankind values structure by noticing our activities must start with some exceptionally harsh or starter introductory presumptions about the presence of the qualities it is endeavoring to gain knowledge. The meaning of the fundamental mammalian worth structure depends on Panksepp and Biven's work, who "arrange the above casual rundown into seven persuasive and enthusiastic components normal to vertebrates: Looking, outrage, dread, desire, mindful, alarm/sadness and play." According to them, Sarma and Hay, at that point, add "neural worth relates," which is some subcortical. That article closes: They contend that what we allude to as human convictions in conversational phrasing can casually deteriorate into (1) mammalian qualities, (2) human idea and (3) human social and social advancement over numerous hundreds of years"

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

Methodology and Results

The study's approach 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 (Meuhlhauser and Helm, 2012). The analyzed published sources are periodicals and non-periodicals. A segment of the data was presented in graphs to explain the presented narratives effectively Fig. 1.

Fig. 1: Flow chart showing how information for the research was collected

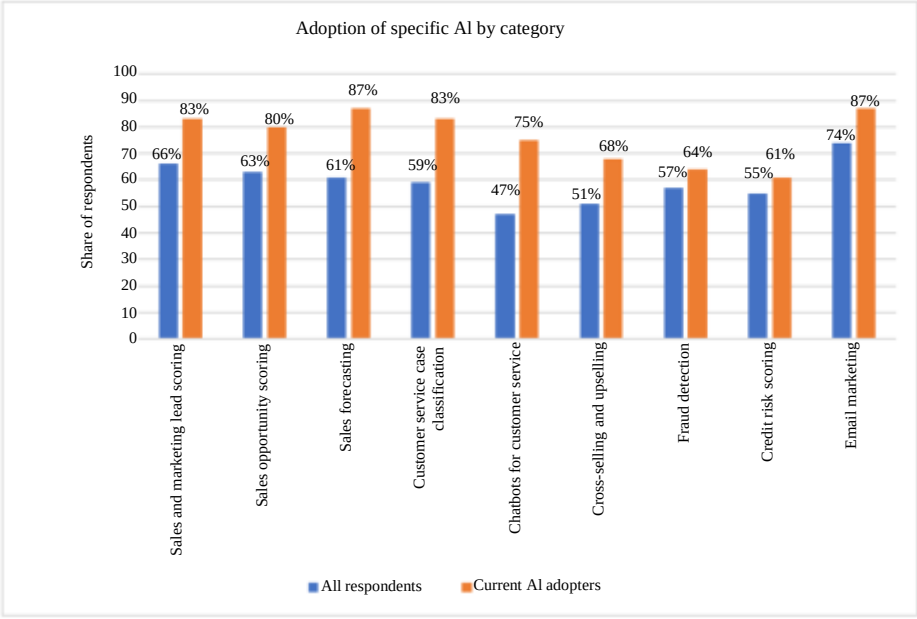

Fig. 2: How various sectors have adopted AI

Table 1: Shows the categories for AI adoption

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	47	75
Cross-selling and up-selling	51	68
Fraud detection	57	64
Credit risk scoring	55	61
Email marketing	74	87

The information provided in table 1 above shows how AI has been adopted in different category. The category with the highest percentage of adoption is sales and marketing with 66%. The second is sales opportunity with 63%, sales for forecasting with 61%, customer service case classification with 59% and chatbots for customers services at 47%.

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.

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 result from evolutionary fitness and can be extracted in the same fashion as Omohundro's AGI fundamental drives from basic evolutionary considerations (Omohundro, 2008). This could describe the most fundamental human drives, such as survival, gender, status-seeking and discovery instinct.

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 *et al*., 2012). In a corresponding article, we will discuss such a hypothesis. We may infer that internal consistency and scientific support and comprehensive literature are 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 2 shows how AI has been adopted in various sectors of the economy. The sectors highlighted require human intelligence for effective operations, thus the need and question for integrating emotive intelligence in AI.

Conclusion

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 supercomputers. Making AGIs as smart as humans will require the customization of AI-based bots to enable real-time monitoring for marketing processes; thus, the AGU frameworks will replicate overly similar intelligence. The 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.

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

Research Limitations

This research paper used secondary published data as the main source of information. The information corrected may not be very accurate as when a primary source is used. Also, the sample used in the secondary data is small and limited. The data for this study was collected from the journal and not included the doctoral and master thesis and dissertations and unpublished sources; therefore, future works would focus on those sources. Moreover, this review study used the English language journals, then, the journals in the other different language can focus on those journals. In addition, this review paper focused on of artificial general intelligence topics; in this regard, further work can be done for the different types of machine learning models.

Acknowledgement

It is my wish to thank all those who played a part towards the completion of this paper. Firsr, I would like to thank my colleagues and fellow authors Ahmed and Alanoud for their unwavering support. If it wasn't for their help with some critical data, this research would not be published. I would also like to thank my professors for providing positive feedback on the appropriate materials to use in the research. To my parents, I thank you for supporting me emotionally and mentally, giving me the hope that this will be a successful project. Despite the challenges faced in the process did writing the paper, it with a fulfilled heart that I thank each of you.

Author’s Contributions

All authors equally contributed in this work.

Ethics

After the publication of this manuscript, there are several ethical issues, which may arise. Ethical issues can arise following cases of plagiarism, lack of etiquette, conflict of interests, and electronic or manual duplication. Ethical integrity of the research will be achieved by ensuring that these issues are avoided.

References

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

Dai, W. (2018). A general model of safety-oriented AI development-LessWrong. https://www.lesswrong.com/posts/idb5Ppp9zghcichJ5/a-general-model-of-safety-oriented-ai-development

Maxwell, J. (2017). "Friendly AI through Ontology Autogeneration" (John Maxwell blog). Https://medium.com/@pwgen/friendly-ai-through-ontology-autogeneration-5d375bf85922

Meuhlhauser, L., & Helm, L. (2012). Intelligence explosion and machine ethics. Singularity Hypotheses: aScientific and Philosophical Assessment, 101-126. https://intelligence.org/files/IE-ME.pdf

Omohundro, S. (2008). The Basic AI Drives. https://selfawaresystems.files.wordpress.com/2008/01/ai_drives_final.pdf

Sarma, G., & Hay, N. (2017). Mammalian value systems. Informatica, 41(3). https://arxiv.org/abs/1607.08289

Sarma, G. P., Hay, N. J., & Safron, A. (2018a, September). AI Safety and Reproducibility: Establishing Robust Foundations for the Neuropsychology of Human Values. In International Conference on Computer Safety, Reliability and Security (pp. 507-512). Springer, Cham. https://doi.org/10.1007/978-3-319-99229-7_45

Sarma, G. P., Safron, A., & Hay, N. J. (2018b). Integrative biological simulation, neuropsychology and AI safety. arXiv preprint arXiv:1811.03493. https://arxiv.org/abs/1811.03493

Vita-More, N. (2020). Wisdomas meta-knowledge: a practical application of artificial general intelligence and neural macrosensing. The Age of Artificial Intelligence: An Exploration, 291.

Worley, G. (2018). Formally Stating the AI Alignment Problem. https://mapandterritory.org/formally-stating-the-ai-alignment-problem-fe7a6e3e5991

Yampolskiy, R. V. (2019). Personal universes: A solution tothe multi-agent value alignment problem. arXiv preprint arXiv:1901.01851. https://arxiv.org/ftp/arxiv/papers/1901/1901.01851.pdf

Yudkowsky, E. (2011, August). Complex value systems in friendly AI. In International Conference on Artificial General Intelligence (pp. 388-393). Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-22887-2_48

Hibbard, B. (2012, December). Avoiding unintended AI behaviors. In International Conference on Artificial General Intelligence (pp. 107-116). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35506-6_12

Muehlhauser, L. (2013, July 16). Roman Yampolskiy on AI Safety Engineering. https://intelligence.org/2013/07/15/roman-interview/?fbclid=IwAR2AOyUc0JEySlgclbwoNcYzL7RzeUnyOFerHqQEgIEWGfuzkGCZQYCIWOg

Sezener, C. E. (2015, July). Inferring human values for safe AGI design. In International Conference on Artificial General Intelligence (pp. 152-155). Springer, Cham. http://agi-conf.org/2015/wp-content/uploads/2015/07/agi15_sezener.pdf

Sotala, K. (2016, March). Defining Human Values for Value Learners. In AAAI Workshop: AI, Ethics and Society. https://intelligence.org/files/DefiningValuesForValueLearners.pdf

Almeida, P., Santos, C., & Farias, J. S. (2020, January). Artificial Intelligence Regulation: A Meta-Framework for Formulation and Governance. In Proceedings of the 53rd Hawaii international conference on system sciences. http://128.171.57.22/handle/10125/64389

Arbital. (2015). Value. Retrieved November 4, 2020. website: https://arbital.com/p/value_alignment_value/