Democracy and Artificial General Intelligence

Conference Paper · January 2024

DOI: 10.54941/ahfe1004960

2 authors, including:

Elina Kontio
Turku University of Applied Sciences
38 PUBLICATIONS **176** CITATIONS

SEE PROFILE

Human Factors, Business Management and Society, Vol. 135, 2024, 330–336
https://doi.org/10.54941/ahfe1004960

Democracy and Artificial General Intelligence

Elina Kontio[1] and Jussi Salmi[1,2]

[1]Turku University of Applied Sciences, Joukahaisenkatu 3, 20520 Turku, Finland
[2]Åbo Akademi University, Tuomiokirkontori 3, 20500 Turku, Finland

ABSTRACT

We may have to soon decide what kind of Artificial General Intelligence (AGI) computers we will build and how they will coexist with humans. Many predictions estimate that artificial intelligence will surpass human intelligence during this century. This poses a risk to humans: computers may cause harm to humans either intentionally or unintentionally. Here we outline a possible democratic society structure that will allow both humans and artificial general intelligence computers to participate peacefully in a common society. There is a potential for conflict between humans and AGIs. AGIs set their own goals which may or may not be compatible with the human society. In human societies conflicts can be avoided through negotiations: all humans have the about the same world view and there is an accepted set of human rights and a framework of international and national legislation. In the worst case, AGIs harm humans either intentionally or unintentionally, or they can deplete the human society of resources. So far, the discussion has been dominated by the view that the AGIs should contain fail-safe mechanisms which prevent conflicts with humans. However, even though this is a logical way of controlling AGIs we feel that the risks can also be handled by using the existing democratic structures in a way that will make it less appealing to AGIs (and humans) to create conflicts. The view of AGIs that we use in this article follows Kantian autonomy where a device sets goals for itself and has urges or drives like humans. These goals may conflict with other actors' goals which leads to a competition for resources. The way of acting and reacting to other entities creates a personality which can differ from AGI to AGI. The personality may not be like a human personality but nevertheless, it is an individual way of behaviour. The Kantian view of autonomy can be criticized because it neglects the social aspect. The AGIs' individual level of autonomy determines how strong is their society and how strongly integrated they would be with the human society. The critic of their Kantian autonomy is valid, and it is here that we wish to intervene. In Kantian tradition, conscious humans have free will which makes them morally responsible. Traditionally we think that computers, like animals lack free will or, perhaps, deep feelings. They do not share human values. They cannot express their internal world like humans. This affects the way that AGIs can be seen as moral actors. Often the problem of constraining AGIs has used a technical approach, placing different checks and designs that will reduce the likelihood of adverse behaviour towards humans. In this article we take another point of view. We will look at the way humans behave towards each other and try to find a way of using the same approaches with AGIs.

Keywords: Democracy, Society, AI, Artificial intelligence, Artificial general intelligence, AGI

INTRODUCTION

Artificial neural network applications have provided the scientific world with valuable new tools ever since the 1980's. Deep learning neural networks have produced astonishing results in e.g. image analysis and classification tasks. Language processing tools have been slower to emerge from the neural network research fields. In 2022 the Transformer network -based chat tools like ChatGPT finally made the development of neural networks visible to a large audience. ChatGPT can process vast data sets from a large variety of subjects to produce text that feels natural to a human. It has even been hinted that the tool has been developed further, so that it can someday solve complex mathematical problems (Tong et al., 2023).

This raises the question whether we are already approaching singularity, the point when AI systems can develop new AI systems better and faster than humans. This would mean the start of a rapid development of AI systems which would lead to an Artificial General Intelligence machine (Azulay, 2019; Dilmegani, 2021). It has been feared that this could lead to a conflict between humans and AGIs because they would compete for the same resources and the AGI system would soon have superior intelligence (Sotala & Yampolskiy; 2014; Yudkowski, 2001; Yudkowski, 2008; Boström, 2014, Salmi, 2022).

In this article we will discuss the problem of coexistence of humans and AGIs. In literature the technical approach has been emphasized: how to design the system architecture so that the AGI is physically not able to harm humans during a possible conflict. Here we adopt a different position: how a society consisting of humans and AGIs can be constructed so that there are incentives for the AGIs to follow commonly accepted rules, like humans do in humans-only societies. We will look at the way humans construct such societies and how they behave towards each other each other and try to find a way of using the same approaches with AGIs.

Humans value conscience and empathy in social relations. Unfriendly behaviour is expected to cause feeling of guilt and a psychologically normal person is likely to avoid that kind of behaviour. An AGI doesn't necessarily have the same kind of system of feelings. It might even be psychopathic. In the USA, while about 1 per cent of population is estimated to be psychopathic, they make about 15–25% of prison population (Hare, 1996).

The rules of a culture's ethics are learned when growing up (Allen et al., 2005; Hall, 2011). When a child grows up, she learns the right kind of behaviour and unwanted habits will be removed through education.

Humans are motivated by a will to satisfy their drives (e.g. hunger, reproduction, improving their status). It is possible to break the society's rules to satisfy one's needs better, but this will lead to punishment. How will the AGI be punished? It could be denied resources or status in the society, making it more difficult to influence others in the future.

MORALITY

Humans and AGIs living as equals in the society requires many things from both parties. They must respect the same rules of ethics. Defining exactly ethical commandments is difficult (Allen et al., 2005). Many ethical human

responses to situations are handled in the subconscious reasoning. Subconscious reactions are fast and don't disturb things in the consciousness, and they presumably handle the cases that are clearer and don't require complex evaluations of reacting.

Ethical rules must be used in certain cases. This makes it easier to agree on measures to take in important decision making, such as medical decision making. In a decision of whether a patient should receive some expensive treatment or not, it would be beneficial to agree on the rules for treatment prioritization beforehand so that they can be discussed.

The humans and AGIs must communicate with one another. Human communication often encounters misunderstanding and unintentional wrong meanings. AGIs must use the same language system with humans but they can also use fast and reliable communication channels of computers to communicate between themselves. But they too can select what they share, and they do not have exact information of each other's memory contents. AGIs will also negotiate, and they may have different views. So, they do not necessarily act as one.

AGI AS A CITIZEN IN A DEMOCRATIC SOCIETY

Should AGIs be thought of as machines or personalities. AGIs are seen as more than normal machines used by humans in that they can themselves take the initiative in a number of tasks. It will form its own personality by making choices. This is different from a normal computer which chooses actions from a narrow set defined by humans and it's the human who has the personality (Salmi, 2022).

On a psychological point of view, we are likely to treat a computer that speaks and acts like a human as a human. Slaves are an interesting analogy. Slaves are somebody's property, and their self-determination is restricted, but yet they are human. Another way of thinking about the limits of free will of the AGIs are Sartre's writings about human free will. According to Jean-Paul Sartre animals just exist (being-in-itself), but humans cannot be determined from the outside (being-for-itself) (Burgat & Freccero, 2015). Maybe the AGIs could be classified as having being-for-themselves as humans do. Sartre says that humans must continuously recreate themselves. AGIs would do so to because they would be able to physically alter themselves.

We argue in this article that AGIs could be controlled in the same way that humans can by making AGIs that have similar ethical characteristics and structures of personality as humans do. Such a machine can be approached in ways like those used to prevent violence and conflict among humans. There are similarities between humans and AGIs, but there are also differences. AGIs can reprogram themselves quickly and more thoroughly than humans can. A human's personality changes slowly, if at all. AGIs can have a more efficient and faster working memory and more processing power. AGIs communicate between themselves faster which may improve their capability of working together.

Human societies come in many forms. In an autocracy there is one supreme leader with absolute power. In a democracy the members of the society each

have one vote and together they select the leaders for a period of time. The democratic institutions set laws and the police and judiciary system control that the laws are obeyed by individuals. An individual or a small group of individuals not obeying the law are unable to resist the police which forces the punishments set by the law. The laws cannot be changed suddenly and there is inertia in the system which guarantees that small changes in opinions don't affect the general structure and dynamics in the society. In many democracies especially minorities are protected even if they don't have many voters.

The democratic institutions also have a role in allocating common resources. This guarantees fairness to everybody in the society. It also makes it difficult for a single person to obtain all the resources.

One of the dystopic views related to omnipotent artificial intelligence assumes that the AGIs are too powerful to be controlled by humans. Here it probably thought that the AGI is one extremely powerful AGI or a group of AGIs that work seamlessly together. It's clear that it is in principle possible to create such a superior AGI. We argue that AGIs can be controlled by making them such that they are equal members of a human-computer society, with the same rights, limits and obligations as humans, even though their capabilities differ from humans.

In a democratic society the most important resources are common, and their division is agreed on by democratic means (Brown & Mobarak, 2004). When there are several AGIs and humans they have to compete for limited resources. In parallel, there is also competition between humans, but it is today regulated by law. The AGIs would also face punishment if they break the law. Could a society consisting of both AGIs and humans control humans and individual AGIs so that they don't break rules? The society can impose punishments. For humans this means fines or imprisonment. For AGIs imprisonment might not be meaningful, but maybe there could be physical limitations imposed on the criminal AGIs.

Human societies are not always very stable, and they can change from democracies to tyrannies. However, the present-day advanced democracies have proved to be relatively stable. This is perhaps due to the fact that even difficult differences in views and goals can be negotiated and there exists means of including every larger group's goals in the common program. Humans don't desire unstable societies because it benefits only very few and, of course, even those in control don't know how long their fortunes will last (Frey & Stutzer, 2000).

Democracy is a process where information flows from the individuals to the deciding organs and vice versa all the time. Different actors in the society have different expectations and roles. Teli et al. (2018) discuss how different expectations can be taken into account in a democracy by making use of a participatory design. The participants make explicit their position in certain issues by selecting positioning cards which best describe their attitude or role in the project. The distribution of the cards shows which are the things that the participants value and, as a result, a common project can be directed according to these values.

Free discussion is a prerequisite that the decisions will be accepted. If some group is silenced, they will feel left out and don't want to be a part of the

society. Therefore, organizing discussion between humans and AGIs would be essential. Hrudka (2020) discusses the way Facebook has changed discussion between humans. Facebook and other social media forums host an endless number of political discussions. Yet they are not just open forums, they have the capability to direct discussion, censor it and amplify certain issues using complex artificial intelligence algorithms. AGIs and humans should be able to take part in the same discussions.

It has even been suggested that the parliamentary authority could be replaced by artificial intelligence (Burgess 2021). In a representative democracy the elected representatives interpret their voter's preferences and act accordingly in the parliament. The representatives could be replaced by algorithms that debate with each other and vote. Their values and goals would be learned from the constituency based on messages from the voters or even by automatic information collecting via internet (Burgess 2021).

RESULTS

In previous sections we have developed the idea of building AGIs that can act as members of a democratic society. This requires that they resemble humans in many ways. They must understand morality and they must be able to communicate their ideas with humans and each other. There would not be one extremely powerful AGI which could dominate humans and other AGIs with superior intelligence and physical powers.

A moral argument can be made for democracy. Such an argument was given by Jean-Jacques Rousseau (1974) in his book The Social Contract from 1762:

"Let us then admit that force does not create right, and that we are obliged to obey only legitimate powers"

Legitimacy then comes from the fact that the inhabitants of a country would vote to discover the common goal. A human or a group of humans has no right to coerce or enslave the rest of the citizens by force. The society must therefore be based on legislation which follows the general will of the people. According to Rousseau, we are only obliged morally to accept orders from a democratic authority. This fits well with the ideas presented in this article. Humans and AGIs would have to agree on this. In the future this could mean an authority that was elected by humans and AGIs.

This is the best case, of course a powerful AGI might emerge which would want total control. It has happened with human dictators many times. But the idea is to try to create a society that is as stable as possible, and which acknowledges the plurality of citizens and yet provides enough added value to each of them to appreciate it. Our argument is that democracy provides that better than alternatives.

We leave open the way how this can be accomplished, but some things are clear:

1. No AGI should be too powerful
2. A voting system for humans and AGIs
3. Forums for communication

4. AGIs should have an understanding of human values and ethics.
5. Legislation which defines the relations between the members of the society and allocation of resources

CONCLUSION

We have discussed the problem of building a society which can incorporate both humans and AGIs. The idea to propose this comes from the uncertainty of guaranteeing the superiority of humans in the future. The artificial intelligence techniques and computer hardware advance so fast, that this may be a problem in a few decades. It may not happen, but it is potentially an existential risk, so it is worth discussing.

Democracy was selected as the model around which such a society could be built. Democracy allows the peaceful coexistence of diverse groups. Every group has means of participating in decision-making and getting their voice heard. This might not mean exactly the same kind of democracy as before because for example ways of communication and the concept of a voter might have to be expanded. If an AI system makes copies of itself, will they be all be eligible voters?

It is clear that the proposed model has also implications on the way AGI systems must be built. They must incorporate such structures that they can participate in a democratic system. These have been left open here, but the necessary features like personality and morality have been analyzed. If AGI systems can be built, I'm sure that these kind of components can also be added to the system. They should be studied well beforehand, so that the philosophical concepts have reached a degree of maturity at the right time and not too late.

REFERENCES

Allen, C., Smit, I. & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics and information technology, pp. 149–55. Springer.

Azulay, D. (2019). When Will We Reach the Singularity? – A Timeline Consensus from AI Researchers. https://emerj.com/ai-future-outlook/when-will-we-reach-the- singularity-a-timeline-consensus-from-ai-researchers/ (fetched 21.1.2021)

Baddeley, A. (2003). Working memory: looking back and looking forward. Nature Reviews Neuroscience 4, pp. 829–839.

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

Brown, D. S. & Mobarak, A. M. (2004). The Transforming Power of Democracy: Regime Type and the Distribution of Electricity. American Political Science Review 103, pp. 1–35.

Burgat, F. & Freccero, Y. (2015). Facing the Animal in Sartre and Levinas. Yale French Studies 127, pp. 172–189.

Burgess, P. (2021). Algorithmic augmentation of democracy: considering whether technology can enhance the concepts of democracy and the rule of law through four hypotheticals. AI & Society.

Dilmegani, C. (21.1.2021). 995 experts opinion: AGI / singularity by 2060. https: //research.aimultiple.com/artificial-general-intelligence-singularity-timing/

Frey, B. S. & Stutzer, A. (2000). Happiness Prospers in Democracy. Journal of Happiness Studies 1, pp. 79–102.
Hall, J. S. (2011). Ethics for Self-Improving Machines: in M. Anderson & S. L. Anderson (Eds.) Part V - Visions for Machine Ethics. Cambridge University Press.
Hare, R. D. (1996). Psychopathy: A clinical construct whose time has come. Criminal Justice and Behavior, 23(1), pp. 25–54.
Hrudka, O. (2020). 'Pretending to favour the public': How Facebook's declared democratising ideals are reversed by its practices. AI & Society.
Knemeyer, D. & Follett, J. (2020). AI, Architecture, and Generative Design. https://towardsdatascience.com/ai-architecture-and-generative-design-e22320828d46 fetched 21.1.2021
Rousseau, J.-J. (1974). The Essential Rousseau: The Social Contract, Discourse on the Origin of Inequality, Discourse on the Arts and Sciences, The Creed of a Savoyard Priest. New York: New American Library.
Salmi, J. (2022) A democratic way of controlling artificial general intelligence. *AI&Society*. https://doi.org/10.1007/s00146-022-01426-x
Sotala, K. & Yampolskiy, R. V (2014). Responses to catastrophic AGI risk: A survey. Physica Scripta, 90(1).
Teli, M., De Angeli, A. & Menéndez-Blanco, M. (2018) The positioning cards: on affect, public design, and the common. AI & Society 33, pp. 125–132.
Tong, A., Dastin, J. & Hu, K. (4.12.2023) OpenAI researchers warned board of AI breakthrough ahead of CEO ouster, sources say. https://www.reuters.com/technology/sam-altmans-ouster-openai-was-precipitated-by- letter-board-about-ai-breakthrough-2023-11-22/
Yudkowsky, E. (2001). Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures, https://intelligence.org/files/CFAI.pdf (fetched 21.1.2021).
Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in global risk: in Boström, N. & Cirkovic, M. M. (Eds.) Global Catastrophic Risks. Oxford University Press, Oxford, UK.