AI & SOCIETY
https://doi.org/10.1007/s00146-022-01426-x

OPEN FORUM

A democratic way of controlling artificial general intelligence

Jussi Salmi[1,2]

Received: 9 March 2021 / Accepted: 28 February 2022

Abstract

The problem of controlling an artificial general intelligence (AGI) has fascinated both scientists and science-fiction writers for centuries. Today that problem is becoming more important because the time when we may have a superhuman intelligence among us is within the foreseeable future. Current average estimates place that moment to before 2060. Some estimates place it as early as 2040, which is quite soon. The arrival of the first AGI might lead to a series of events that we have not seen before: rapid development of an even more powerful AGI developed by the AGIs themselves. This has wide-ranging implications to the society and therefore it is something that must be studied well before it happens. In this paper we will discuss the problem of limiting the risks posed by the advent of AGIs. In a thought experiment, we propose an AGI which has enough human-like properties to act in a democratic society, while still retaining its essential artificial general intelligence properties. We discuss ways of arranging the co-existence of humans and such AGIs using a democratic system of coordination and coexistence. If considered a success, such a system could be used to manage a society consisting of both AGIs and humans. The democratic system where each member of the society is represented in the highest level of decision-making guarantees that even minorities would be able to have their voices heard. The unpredictability of the AGI era makes it necessary to consider the possibility that a population of autonomous AGIs could make us humans into a minority.

Keywords Artificial intelligence · Artificial general intelligence · Democracy

1 Introduction

Throughout history, human society has only seen humans as having an active role. Animals have and other non-humans have not had the same rights and responsibilities as humans (Wojtczak 2022). They do not understand the law and they are under the responsibility of humans. Lately, attention has been given to the influence AI systems have in humans. For example, social media algorithms pursue political goals without being constantly operated by a human (Hrudka 2020). The speed at which the AI algorithms develop and the resources that are being poured into their development has led many to believe, that the day when the artificial general intelligence machines start to make their own decisions without human initiatives will be reached at some point. The estimates by the majority of top AI researchers currently predict that computers will be able to make these advances during this century (Azulay 2019; Dilmegani 2021).

Why is this worth worrying about? It is likely that humans would be inferior in intelligence to these AGIs. For this and other reasons, there is a potential for conflict between humans and AGIs. Further, AGIs are different from other machines in that they would have autonomy that make them independent of humans. They might even set their own goals and have the means to pursue them. Today, specialist systems are not independent actors except in their narrow areas of use. So, could humans or the society of humans and AGIs influence the way that autonomous AGIs set their goals? What is the worst-case scenario and how likely does it seem to be? (Sotala and Yampolskiy 2014; Yudkowski 2001; Boström 2014).

In the worst cases, AGIs would harm or kill humans either intentionally or unintentionally. Unintentional harming would happen e.g., if AGIs misinterpret what humans want or they accidentally created circumstances which were not beneficial for humans. It is difficult to estimate how likely this would be, but it is clearly a possibility and, therefore, it is wise to think about ways to protect human lives/

✉ Jussi Salmi
salmenjussi@gmail.com

[1] Research IT, University of Turku, Turku, Finland

[2] Department of Computer Science, Åbo Akademi University, Turku, Finland

Published online: 18 March 2022

values. The biggest factor behind such fears is that AGIs would not be able to understand or do not care about human motivations or values.

A comprehensive review of the problem of controlling AGIs was written by Sotala and Yampolskiy (2014). They argue that the main problem is that of the unpredictable nature of AGIs. They may break the status quo quite fast and in so doing, surprise humans. Eliezer Yudkowsky has propagated the idea of a friendly AI (Yudkowski 2001, 2008). Friendliness would be guaranteed using clever fail-safe mechanisms while programming the AI, if that is possible. Nick Boström described an unfriendly AI and its societal implications in Superintelligence: Paths, dangers, strategies (Boström 2014).

This article is organized as a thought experiment. First, we imagine an AGI which is a good citizen in a democratic society (Sects. 2.1 and 2.2). To present this in a meaningful way, it is necessary to dive more deeply into what is needed from an artificial intelligence being to be more on par with humans in terms of autonomy and to be socially less machine-like in human-AGI-interaction. Then we discuss how such an AGI could function in a democratic society consisting of both AGIs and humans (Sect. 2.3). In Sects. 3 and 4, we discuss the implications of having such a society.

2 Materials and methods

2.1 Artificial general intelligence in human world

The view of AGI we have in this thought experiment is such that it resembles humans a lot. It acts as an agent that has a commitment to its own motivations and values. Keeping these intact, it can set its own goals and have a complex internal life that enables it to be perceived as a person. Accordingly, psychological concepts can be used to analyse its behaviour. It may be seen as possessing a Kantian autonomy in a sense where, like a human individual, the device has human-like values, urges and goals. This is by no means easy, and breakthroughs are required to realize this not only in technology but also in philosophy and psychology. Nevertheless, this is an assumption in this thought experiment.

Given such a design, the AGI will seek to satisfy these desires and, thus, come into conflict with humans or machines with which it is in competition for resources but also has cooperation with. Current specialist systems are to a large degree controlled by their human operators and, for this reason, do not face independent questions of morality or make decisions of other than limited scope. Autonomic decision-making will require choices between alternative actions. Accordingly, as in in humans, these are likely to constitute some building blocks of the personality. This kind of architecture is needed in this thought experiment to make them good citizens in a democratic society. A citizen AGI takes part in governing the society by electing representatives and acting as a representative. Thus, it must understand the goals and actions of other members of the society, and it must be able to formulate its own goals in a way that can be turned into democratic action. Its values may change over time when it interacts with others. So, in this section and in Sect. 2.2, we will analyse how humans work and at the same time, in what way should the AGIs work to cooperate with humans. Humans are unable to operate like computers but perhaps, for the sake of the argument, the AGIs can mimic or at least understand human behaviour for the purpose of cooperation even though they can purely technically most likely perfectly well operate without that.

The Kantian view of autonomy can be criticized because it neglects the socially derived character of human beings. Humans live through and with each other to fulfil their needs. The same may or may not apply to AGIs. Their individual level of autonomy and cooperation determines how strong is their society and how strongly integrated they would be with the human society.

To understand what this requires from AGIs we must first discuss humans. The current democratic society works in ways that take into account the capabilities and shortcomings of humans. So, first we will consider the question how and why humans act as they do. In sharing the society with humans, the AGIs must understand humans' motivations. If they don't, they judge humans' actions wrong. Having these capabilities AGIs would be better accepted by humans. After all, in a democratic society, an actor with unwanted or poorly understood behaviour will face difficulties in gaining support for his/her goals. When we talk about human minds, we cannot neglect the importance of consciousness. On a classic view, consciousness consists of sensory data, knowledge of self and short-term memory (Baddeley 2003). Given consciousness we use internal (and external) talk to plan and assess action and to direct our attention. Consciousness exists because our physiology makes it possible, and it helps our thinking. Humans only see conscious beings as morally responsible and equal. A person is morally responsible only for things that he/she could have done differently (Klein 2005). To see AGIs as equal to humans their actions and "thoughts" must be understandable to humans. The AGIs must be seen as conscious enough to be held accountable and appreciated for their actions.

The role of consciousness in handling emotions is central to decision-making (Tsuchiya and Adolphs 2007). All normally functioning humans have feelings of guilt after breaking the rules and behaving badly. The feeling of conscience then enters consciousness. Internal talk can shape self-blame. A reprehensible person is one that lacks conscience, or the knowledge and feeling that one has done wrong. Because humans have a knowledge of self and conscience the feeling does not pass quickly. These feelings make us less likely to repeat the action. Humans fear an unfriendly AGI because it does not necessarily have a conscience. It might resemble a psychopath. Psychopathy is not just a human phenomenon. In AGIs it would mean not respecting other's freedom of choice and being unempathetic. This is a central thing in how an AGI could be a threat to humans. It must have its own version of sensitivity and empathy in conflicts. An AGI without empathy would probably not feel bad about the suffering of others when they are e.g. abused. Humans like these are judged to be dangerous. With an AGI these unwanted personal traits might combined with immerse strength work in ways that would pose an existential threat. 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).

Humans strive to satisfy their drives (e.g. hunger, reproduction, improving their status). Humans sometimes satisfy their needs more easily using ethically forbidden means, even crimes. Sigmund Freud (1999) described three levels in human psyche. The basic drives (id) that unconsciously causes the human to seek pleasure. The ego controls the basic drives and finds a rational and efficient way to fulfil them. The super-ego controls that the lower levels (ego and id) don't engage in immoral action. The fear is that the AGI does not have either feelings or the super-ego which makes the AGI follow rules agreed on by the different actors in the AGIs universe. Traditionally, it is thought that controlling the AGI can then require external constraints on the AGI or more refined inner super-ego structure which causes the AGI not to seek goals that are harmful for humans. As an example, an AGI can have "drives" or goal calculation functions that get maximum points without using too many of the common resources (Shulman 2010).

Computers don't have the same kind of limitations as humans. They don't need consciousness like humans to have a complex personality. What drives an AGI? Biological systems like humans and animals need, above all, to find enough food for successful reproduction. A computer needs electricity which is available from the electricity grid. Extra electricity beyond its needs does not interest it and the quality of electricity is always the same. Reproduction matters less because a computer has an unlimited lifetime and lacks evolutionary pressures.

2.2 Morality

The goal that humans and AGIs would live together in a democracy would require both kinds of system to share similar values and understanding of moral justification. So, can we implement human democratic or moral values in an AGI? It seems difficult to define an ethical list of commandments (Allen et al. 2005). Humans use feelings as well as both subconscious and conscious mechanisms to determine an ethically correct way to react. Subconscious reactions are fast, and they presumably handle the cases that are clearer and don't require complex evaluations of several potential ways of reacting.

For example, if one has a chance to jump the queue in a shop you know it is wrong without making the matter explicit. But if one finds a 50 euro note in the shop floor many would consider whether to take the money to the cashier or to put it in your wallet. It requires more complex consideration; it is brought into the consciousness. It seems to be very hard to formulate rules for such behaviour. Such a rule might say that "if you find things that don't belong to you take them to some official". But if you find only a 1-euro coin on the floor it wouldn't be ethically wrong in most people's minds to just take it. But then again, if you just heard the sound of a coin dropping to the and an elderly woman is standing close by most people would ask her whether it's hers and not just pocket the money. So, the set of ethical rules that people follow is practically infinite and it is impossible to formulate the rules in a top-down manner (Allen et al. 2005).

Nevertheless, there are situations when clearly defined rules are used. Perhaps, therefore, a hybrid approach to morality could be developed. There would be a moral "sense" which defines the way simple situations are solved without relying on a potentially endless set of logical rules. On top of those, more complex situations could be handled by more explicit rules. This makes it easier to agree on measures to take in certain situations, such as medical decision-making. When deciding 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.

Humans learn the ethical rules in childhood from examples and they can interpret them in a wider context as general rules. Their neural network in the brain self-organizes through development, trial and error and feedback during many years. The information processing in the brain at low level does not include symbolic processing. Accordingly, there are no clearly textually defined rules.

In humans, complex ethical rules are handled in consciousness (Milner and Goodale 1995). Whereas the non-conscious cannot handle conceptually complex, consciousness enables internal talk where concepts can be easily cut into pieces and handled piece by piece in a slow and error-prone manner. In so doing, a person (or a neural executive structure) uses working memory to draw on memories, feelings, emotions and sensory input. The neural structure of consciousness is largely unknown. For AGIs it may not be needed at all in that there are no grounds to think that complex reasoning (or rule-following) depends on consciousness. Or perhaps, for AGIs, consciousness could be omnipresent especially if, as Baddeley (2003) and others argue, consciousness is highly dependent on the fast use of working memory. The working memory of an AGI can be vast, and unlike humans, it can hold the whole idea and all of its details in working memory at the same time.

A central part of a society is the ability to do things together and communicate with each other. Without it, cooperation is impossible. The communication between humans happens via languages. Although the sender and receiver may know the standard meanings of the words, variations and differences between people cause problems in the communication. Misunderstanding is common due to the receiver selecting interpretations that may not match the sender's because, in fact, most communication is ambiguous. AGIs won't necessarily face these issues. If they follow the same standard architecture, with similar data structures or neural architecture, they can transfer the contents of working memory directly as they are represented in their programs' data structures as one big data structure. They can thus transfer whole working memory contents exactly and in a subsymbolic way. But they too can select what they share and they do not have exact information of each other's memory contents. An AGI may choose to mislead others or withhold information to gain something using game theoretical tactics and they too have to go through negotiations to find common ground. This is important because it means that the AGIs may find it impossible or not advantageous to build a unified front against humans if some of them think that it is more advantageous to ally with some humans in negotiations about common issues.

AGIs are seen as more than normal machines used by humans in that they can themselves take the initiative in a number of tasks. E.g. if a house robot has washed the dishes and mowed the lawn as you requested, it is not yet an artificial general intelligence. But if it has decided by itself to build a doghouse for the family's pet and it can be called an AGI. The choices it makes reflect the options from which it chooses. 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.

A human can doubt whether another human possesses consciousness, feelings or an internal narrative. Since we only see what another person says or how he behaves, we can be sceptical about their use of consciousness. The same applies to artificial general intelligence computers. We don't see the state of its program execution, but we observe its speech and behaviour. If a computer speaks and acts like a human, surely we will behave towards it more or less as we would behave with a human person. Perhaps at first, we will treat it like a slave so that we don't have to react to its wishes, and we treat it as somebody's property. Or perhaps if the AGI starts to show simulate feelings towards us we will start to care about its 'feelings'. According to Jean-Paul Sartre animals just exist (being-in-itself), but humans cannot be determined from the outside (being-for-itself). (Burgat and Freccero 2015). One can argue that the AGI, unlike current specialist systems could perhaps be classified as having being-for-themselves as humans do. Sartre says that humans must continuously recreate themselves. AGIs would do so too in very concrete ways because they would be able to physically alter themselves.

2.3 AGI as a citizen in a democratic society

Currently, the AGIs are seen as tools and property of humans, like slaves were earlier in human societies. The upper class were able to keep the slaves oppressed mostly through a monopoly of violence. Slavery ended when it became more profitable to hire capable workers than to force badly motivated and uneducated slaves to work and because others began to argue that slavery was unjust. In a moral reset, they reinterpreted the ambiguous claim that every person has a value by itself by including slaves (who were previously non-persons). On these grounds, it seemed wrong to treat a person with human rights as a property like a machine. If we think that AGIs are human-like personalities with a free will, is it wrong to see them as property of a human-like slaves once were? Will the human powers grant them human rights as well? What implications does this have to using them for performing tasks in the economy and industry?

Above, we emphasised that there are similarities between humans and AGIs There are differences as well, including (1) AGIs don't need feelings in the same way humans do, (2) communication between AGIs is more direct, (3) AGIs have more efficient working memory whereas humans have a narrower consciousness, (4) AGIs can reprogram themselves in an instant, changing their design and physical capabilities. What are the implications of how dissimilarities contribute to a possible existential threat to human life?

In this thought experimented, we assumed that the proposed AGI would be a good citizen and it would take part in democratic decision-making. This would be incompatible with at least the ability of the AGIs to reprogram and develop themselves very fast. It would be very risky to build alliances in elections with partners that are totally untrustworthy—who can change their mind in an instant. Therefore, for the sake of the argument we presume that they are more persistent in their opinions. But how realistic is that? From humans we know that a person loses his ability to influence others if he changes his/her mind constantly about important things, even though doing so may be justified and sometimes a good strategy in decisions concerning only him/herself. There is no reason why this wouldn't be the same even with AGIs. Thus, the argument could be made that it might make it more difficult for AGIs to obtain their objectives if they practice a strategy of constant reprogramming of their political thinking.

Next we will discuss the implications of these kind of AGIs to the society. Human societies are organised in many ways. In an autocracy, there is one supreme leader with absolute power. In a democracy, by contrast, the members of the society each have one vote and together they select the leaders for a period of time. Individual members of a democratic society are obliged to obey the democratically enacted laws by way of punishments and finally the violence monopoly of the governing body. An individual or a small group of individuals not obeying the law are unable to resist the police or army which forces the punishments set by the law. In a democracy a larger group of dissidents can obtain power at times; however, it is usually difficult to get enough support for a radical change of law and there is inertia in the system which makes it very difficult to thoroughly change things fast.

For democracy to work, it has to be accepted by people who willingly take part in electing democratic institutions. If an individual suddenly becomes very rich and powerful, would he still be willing to be a part of the democracy? Most likely yes. The reason for this is that humans usually adapt to their society's values. As argued above, humans are typically unwilling to hurt other humans even if it is possible because of their conscience and system of feelings. Society can be valuable also to persons who can guarantee their own prosperity and security by making use of their superior resources.

In a human society many people want to obtain as many resources as possible and, for this reason, there are rules for limiting the use of resources. This guarantees fairness to everybody in the society and, in part, makes it difficult for a single person to obtain all the resources. The rules for division of resources enable most people to live at a historically specific level of prosperity.

In a democratic society, there is an agreement that the most important resources are common, and their division is agreed on by democratic means (Brown and Mobarak 2004). When there are several AGIs, they must compete for limited resources. In parallel, there is also competition between humans, but it is today regulated by law. Also AGIs would have to obey rules for allocating resources to keep the competition fair. 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 such as reduced electricity use or giving up some more resources for an AGI. For an AGI it would be beneficial to accept the punishments because it would receive more serious punishments if it protested the punishments.

There is often instability in human societies and wars can move societies from a democracy to tyranny (and back again). Artificial general intelligence machines might also have such aspirations; however, but if there were a heterogenic group of AGIs they might not form a unified front against others. But there are no insurmountable guarantees about this. Not doing so would be based on not seeing humans as a threat to counter and the inability to find a common ground between all the AGIs. Similarly, in a human society a large group of strong and resourceful individuals can at any time seize power and rewrite the rules to their advantage; however, this doesn't often happen in societies whose structures are stable. Humans don't in general 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 and Stutzer 2000).

Whereas a society with a single AGI could decide on a policy that depends on an enormous use of resources that was harmful to humans, a democratic group of AGIs would have to negotiate within themselves and humans on the appropriate policies. This is because those resources would be controlled by the society. This could mean tight control of communication and electricity networks and other critical infrastructure. It is still possible that some extremist group of AGIs could seize these resources. This could be made difficult by making it harder to change the rules. It is important to note that there cannot be a completely flawless system with as powerful actors as AGIs.

Democracy does not only mean electing new leaders from time to time. It means a deeper commitment to cooperation in the society. Different actors 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 by selecting positioning cards which can 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.

Public discussion is a prerequisite for making informed choices as a democratic citizen. For inter-AGI communication, this discussion will be different from humans, but nevertheless they will also have differences of opinions due to differing needs and experiences among AGIS, even though the needs will change over time. Even for them, discussion may take time before a common view on some discussion item is found and politics will undoubtedly play a part of that discussion.

Hrudka (2020) discusses the way Facebook has changed this discussion for humans. Facebook and other social media giants 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 techniques. AGIs and humans need to be able to take part in the same discussions. This will require appropriate channels for the discussion. These channels must be controlled, but who will control them? Now the control is given to the owner of the forum because the forums of discussion are not seen as independent actors. With AGIs this must be guaranteed better because impartiality must be guaranteed and AGIs cannot be seen as impartial. The controller of the forum can be an AGI having a political view in the future. The controller has huge non-democratic influence in the society. Yet these forums can be very valuable meeting points in a future society consisting of humans and AGIs.

Modern discussion of democracy has even 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. They could be replaced by algorithms that debate with each other and vote. Each algorithm instance would debate about the proposals and try to influence the outcome. It's values and goals would be learned from the constituency from messages from the voters or even by automatic information collecting via internet. (Burgess 2021).

In this section, we have discussed what would a society with AGIs as citizens look like. Further to this, we proposed that some sort of capability to engage into moral thinking is required to construct a human-AGI society, because humans need that as a way of guaranteeing that the society is based on some commonly agreed rules that are written as laws and codified as a certain kind of behaviour. A totally amoral AGI would not be accepted any more than an amoral human being, because it in cannot be relied on to respect contracts or otherwise act as a reliable member of the society. A society must have structures that enable citizens to co-exist. For the coexistence of humans and AGIs, a developed and egalitarian society is to be preferred. In this section, we discuss the idea that democracy is an efficient way of organizing a society.

A moral argument can also 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 sovereignty, consisting of all the people that live in the country's area. A human or a group of humans has no right to coerce or enslave the rest of the citizens, according to Rousseau. The society must therefore be based on legislation which follows the general will of the people. In principle, the general will can be discovered by voting. According to Rousseau, we are only obliged morally to accept orders from a democratic authority. In the future this could mean an authority that was elected by humans and AGIs. This fits well with a moral foundation that can be shared by humans and AGIs.

An AGI can of course reprogram itself to not use any moral and it could then disobey any laws and refuse to be punished. But because the critical resources would still be governed by the democratic society there would be a strong incentive not to do so and to be able to influence the society this could be counterproductive. This democratic model requires that there would a large group of AGIs. A single omnipotent being would be more difficult to control. But will the AGIs respect the laws? There will always be the possibility that they won't, and it could be bad for humans. Even the solely human societies have been close to be completely destroyed by war due to power struggles. AGIs are a powerful class of actors in a society, and they may be impossible to integrate in the society.

3 Discussion

In this article, we have discussed the problem of whether humans and artificial general intelligence machines could coexist peacefully. Although the problem is not yet actual, it seems to be important to circulate ideas about possible strategies in building the AGIs. In a decade or two, may be a time when regulation is needed and it is important that discussion should mature before then. Accordingly, we have analysed some basic concepts and picked up work by researchers who have suggested solutions. Currently we already have specialist AI systems that require ethical consideration. Medical specialist systems make recommendations for the treatment of patients considering the price of treatments, autonomous cars make many decisions which have ethical dimensions, social media algorithms choose content to show to the user. The ethics research has not yet matured. The discussions of today will only show their relevance in a decade or two.

We started this article with describing the risk that AGIs might pose. They are very powerful and fast, and they may have brilliant minds but may completely lack morality. Humans may not be able to control them because they may show a new kind of autonomy. Traditionally most of the discussion has concentrated on one AGI and drafted programmatic structures that try to limit the actions of that one AGI. I have not discussed the challenges of building such an AGI. In parallel with that discussion, we must also pay attention to the issue of what kind of AGI we should build and that is what I have done in this article.

We do not know yet what the artificial general intelligence computers will be like. Perhaps there will be an easy solution and some computational architecture will guarantee the friendliness of the AGIs. Or perhaps they will never be built—computational power may not suffice to build an intelligent computer. We must still know how to program intelligence and many issues about intelligent behaviour remain unclear. But we must prepare also for the case that there will be a conflict between humans and AGIs and we must think about ways to solve that conflict.

In this article, we have discussed the view that AGIs can be controlled by including them in a democratic system. The idea is to design a society which can hold both AGIs and humans. The society with AGIs doesn't have to be large and could be a small bubble inside a larger society. The basic idea is that in a society like democracy with mutual interests as the driving factor it would be easier to harmonise the interests of the different members of the society than in a strictly hierarchical societal structure where there is constant struggle for the top posts and the strongest wins a very powerful authority.

This article is about whether it is possible to build an AGI with a morality component and that democracy is a possible way of organizing a society with humans and AGIs, at least in theory. We also discussed whether a moral AGI can be a good citizen in a democratic society. A democratic society is based on a social contract, all parties wish to be parts of that society because inside it they can best fulfil their own goals.

Further problems include the fact that at first, the AGIs would be a small minority. Later on, humans might be a minority. This would require legislation that guarantees the rights of such minorities. A small minority has only a small influence in elections so they may require extra protection in legislature and democratic institutions. We do not know what kind of AGIs there could be. What is the exact distinction between a simpler artificial intelligence and artificial general intelligence computer? What if an AGI makes a million copies of itself will all of these copies have equal rights to vote? Should the right of an AGI to make exact copies (with also the contents of mind copied) of itself be limited somehow?

Perhaps a central question is what it means for an AGI to be a member of a democratic society. What does their autonomy consist of when part of that autonomy must be given away to accommodate for other society's members' needs? These are things that must be discussed in the future.

Funding Open Access funding provided by University of Turku (UTU) including Turku University Central Hospital.

References

Allen C, Smit I, Wallach W (2005) Artificial morality: Top–down, bottom–up, and hybrid approaches. Ethics and information technology. Springer, pp 149–155

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 Jan 2021

Baddeley A (2003) Working memory: looking back and looking forward. Nat Rev Neurosci 4:829–839

Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press

Brown DS, Mobarak AM (2004) The transforming power of democracy: regime type and the distribution of electricity. Am Polit Sci Rev 103:1–35

Burgat F, Freccero Y (2015) Facing the Animal in Sartre and Levinas. Yale Fr Stud 127:172–189

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 Soc 37:97–112

Dilmegani C (2021) 995 experts opinion: AGI/singularity by 2060. https://research.aimultiple.com/artificial-general-intelligence-singularity-timing/. Fetched 21 Jan 2021

Freud S (1999) In Strachey J (ed) The standard edition of the complete psychological works of Sigmund Freud, vol XIX

Frey BS, Stutzer A (2000) Happiness prospers in democracy. J Happiness Stud 1:79–102

Hare RD (1996) Psychopathy: a clinical construct whose time has come. Crim Justice Behav 23(1):25–54

Hrudka O (2020) 'Pretending to favour the public': how Facebook's declared democratising ideals are reversed by its practices. AI Soc. https://doi.org/10.1007/s00146-020-01106-8

Klein M (2005) Responsibility. In: Honderich T (ed) Oxford companion to philosophy. Oxford University Press

Milner AD, Goodale MA (1995) The visual brain in action. Oxford University Press

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 American Library, New York

Shulman C (2010) Omohundro's basic AI drives and catastrophic risks. http://intelligence.org/files/BasicAIDrives.pdf. Fetched 21 Jan 2021

Sotala K, Yampolskiy RV (2014) Responses to catastrophic AGI risk: a survey. Phys Scr 90(1):018001

Teli M, De Angeli A, Menéndez-Blanco M (2018) The positioning cards: on affect, public design, and the common. AI Soc 33:125–132

Tsuchiya N, Adolphs R (2007) Emotion and consciousness. Trends Cogn Sci 11(4):158–167

Wojtczak S (2022) Endowing artificial intelligence with legal subjectivity. AI Soc 37:205–213

Yudkowsky E (2001) Creating friendly AI 1.0: the analysis and design of benevolent goal architectures. https://intelligence.org/files/CFAI.pdf. Fetched 21 Jan 2021

Yudkowsky E (2008) Artificial intelligence as a positive and negative factor in global risk. In: Boström N, Cirkovic MM (eds) Global catastrophic risks. Oxford University Press, Oxford