FILOZOFIA
Volume 79, Number 5, 2024
DOI: https://doi.org/10.31577/filozofia.2024.79.5.3

Can and Should Language Models Act Politically? Hannah Arendt's Theory of Action in Comparison with Generative AI

LUKAS OHLY, Fachbereich Evangelische Theologie, Goethe-Universität Frankfurt am Main, Germany

OHLY, L.: Can and Should Language Models Act Politically? Hannah Arendt's Theory of Action in Comparison with Generative AI
FILOZOFIA, 79, 2024, No 5, pp. 501 – 513

At the heart of the debate about generative or autonomous AI is the concern that it could take action away from humans. Hannah Arendt, who gave the concept of action a political profile and opposed the automation tendencies of modernity, provides a basis for assessing fears that humans will cede their political action to AI. After an introduction to Arendt's political theory of action, the paper examines whether AI can and should act. The focus is on linguistic models.

Keywords: Hannah Arendt – AI-text-generators – acting – democracy – freedom

I. Why AI Challenges Our Political Rights?

Underlying many of the ethical issues surrounding AI are fears about political participation rights. When autonomous cars or care robots become capable of making decisions, this increase in automation competes with the autonomy of members of society. If generative AI obscures the authorship of scientific or artistic works by transforming them into data and recombining them, claims to validity can no longer be traced and thus lose their social relevance. Without the ability to attribute statements, individual rights lose their social function or can be reduced to a mere form of freedom of movement. Freedom rights threaten to be transferred to networks (Harari 2017, 446).

Many ethical questions about AI relate to its similarity to humans. It can make decisions, speak, write, create works of art and relieve humans of cognitive tasks. In contrast, AI poses a political challenge to humans because of its fundamental *dissimilarity*. I will show that AI cannot act politically, but that it threatens to occupy the political space. This requires a political theory that significantly profiles the concept of action.

In this article, I examine the political consequences of the application of AI using Hannah Arendt's concept of action. Arendt is suitable for this purpose because, firstly, she distinguishes her concept of action from other human activities. This also leads to a clear definition of political space. Secondly, Arendt recognized the dangers of automation for the political sphere more than half a century ago. Even if Arendt's concept of politics is sometimes rejected as idealistic, her scenario can be used to identify the dangers of AI "gaining" a political "voice."

First, I reconstruct Arendt's theory of action, which is a political theory of action. In a second step, I show why AI cannot act in Arendt's sense. Nevertheless, it does have an impact on political communication processes that can restrict people's political freedom, which I explain in the final section.

II. Arendt's Theory of Action

According to Hannah Arendt, action is the only basic activity that people can carry out without the mediation of material things and only together in plurality (Arendt 1998, 7). In my view, examples of this could be the negotiation of common interests or strikes. What supports Arendt's negotiation is her observation that speaking is an essential part of acting: speaking and acting distinguish people from each other and thus form a plurality (Arendt 1998, 176). What supports the strike as a paradigmatic example of action is Arendt's reference to the revolution as the birth of freedom (Arendt 1977, 29, 31).

This brief introduction already reveals the peculiarities of Arendt's action:

1. Action is political. When people act together, they create a public space (Arendt 1998, 57).
2. For this reason, action is also linked to freedom: The freedom claimed in action is a political freedom (Arendt 1978, 200). The terms action and freedom are interchangeable.
3. Action brings something new into the world (Arendt 1977, 34). It interrupts existing causal chains and begins a new causal series (Arendt 1977, 206).

Arendt associates action so strongly with being human that she describes human beings themselves as the beginning (Arendt 1998, 177). At the same time, in her depiction of the new, she recognizes the divine ability to create something out of nothing (Arendt 1977, 206), and transfers it to human beings.

There are two reasons for this: First, action takes place without material mediation and therefore does not even have to resort to causal chains that determine material processes. Second, man is not an isolated individual, but a plurality of individuals who differ from one another in their *common* actions. The beginning, which is the human being, is not an absolute beginning like God's creation out of nothing, but occurs among people who have already begun. In the beginning of something individually new, the difference from existing people also begins. A beginning needs a community in order to *exist*, i.e. to have the ability to persist. And a community of doers *exists* in the beginning because the beginning creates the difference that constitutes human plurality. In order to persist, plurality must remain in the beginning.

Arendt has thus highlighted a significant difference from other human activities. But is this difference valid? Is there really such an action that emerges as something new without material mediation in a community of self-differentiating individuals? One could argue that without material mediation we cannot communicate, i.e. we can neither speak nor act: On the telephone, we would not be speaking in Arendt's sense, and even the sound waves necessary for speaking would have to be excluded. When we act, we would not be allowed to make decisions that have material consequences – for example, political redistribution using tax money would not be acting in Arendt's sense. This political freedom would be completely divorced from any needs that corporal and vulnerable people might have (Habermas 1994, 220; Pitkin 1994, 271).

I therefore understand Arendt's juxtaposition of action and materiality as a categorical opposition. Categorical opposites are not mutually exclusive, but can apply to the same situation – as is the case here: If people need sound waves in order to speak, this does not mean that they explicitly focus on sound waves in order to speak. The mouth is not a tool for them to "produce" words, but they do not usually focus on their mouths. Only when they make a promise or are prevented from speaking do they become aware of their mouth – but at that moment they are prevented from speaking.

The same applies to other actions, which for Arendt are political: We grant each other mutual freedom by jointly granting each other the difference that we are. Arendt linked action so closely to the word that speechless action is almost impossible (Arendt 1998, 179). At the latest when someone asks: "Why are you doing this?" binds an acting person to the word. Typical examples of action for Arendt are forgiveness (Arendt 1998, 241), creating

freedom, creating spaces of freedom to move freely (Arendt 1977, 275) – to start something new at all (Arendt 1998, 177).

Materiality therefore remains in the background of action. Even though it is a prerequisite for people to be able to talk to each other and move freely, they do not communicate by materiality. The difference becomes clear when compared to manufacturing: Anyone who makes something must take the material into account, otherwise the manufacturing process would be completely random. When we act, on the other hand, materiality recedes so much that it seems not to be there at all, as if we were in "direct" exchange with each other.

In tax policy, on the other hand, it seems to me that action on the one hand and production or work on the other are related to the same situation in a categorical opposition. When a country uses taxes to redistribute income between social milieus, the material provision of members of society is of course central. The distribution of tax money is not an action. But the question, how to distribute, is an action, namely in parliamentary negotiation processes and voting. Of course, voting also requires material: Members have to raise their hands or fill in a ballot. But the material does not change the authority of the result of the vote. When parliamentarians agree to a secret ballot, it is not because they believe that the material of the ballot affects the outcome. Rather, they want the freedom of all delegates to be respected.

The distribution of tax money, on the other hand, takes place outside parliament – and therefore outside the political arena. Anyone who has ever had to file a tax return knows how much technology is involved, from the tax program to the various tax forms: Whoever fills out the wrong form, changes the tax rate automatically. Taxes are political, but that doesn't mean that every step of the tax calculation process is political. "The liberation from the curse of poverty would come about through electrification, but the rise of freedom through a new form of government, the *soviets*" (Arendt 1977, 65f).

The meaning of action, on the other hand, lies in direct interpersonal interaction – ignoring material conditions. This meaning is the recognition among members of society that they are different and have the freedom to be different: "The equality attending the public realm is necessarily an equality of unequals" (Arendt 1998, 215).

I consider action, as Arendt points out, to be a real phenomenon. At the same time, it has already become clear how few examples of action there are: Forgiving, negotiating, recognizing each other as free persons, and above all, speaking. When shopping at the checkout, payment processes can now be

automated by machines that register the goods and collect them. But there is no action. When I face a human cashier, the same processes take place, but moreover we perform a mutual recognition as free persons. But this fact has nothing to do with the payment process. The contrast between acting and other activities is categorical: several activities can come together in the same situation.

At the same time, I believe that Arendt's paradox is true, that the meaning of action does not precede it, but arises along with it. It arises when common opinions are formed. No one can decide alone to share a common opinion with others. Even if I "join" an opinion, my decision to join must be based on the fact that I already share that opinion with others. So in my decision I have already missed the moment when we have begun to agree on an opinion. This political space of mutual recognition thus arises "out of nothing." In retrospect, I can justify why I agree with this opinion, and Arendt points out that this retrospective look is dependent on another categorically different practice, that is judging of a spectator who has an overview of the whole (Arendt 1992, 44). But the justification for agreement could not have caused my agreement because it came too late: The actors and the spectators are originally different persons (Arendt 1992, 15).

III. Can AI Act?

The partial moments of action described above already rule out the possibility that AI can act: It cannot "share" opinions, nor can it break out of existing causal chains to start something new, simply because whatever it generates seemingly new is an immediate effect of its training data and its program. Its seeming "novelty" is not embedded in an immediate social sphere which characterizes novelty in Arendt's theory of action. Thus AI cannot have free recognition "out of nothing." It is not political insofar as it does not reciprocally grant freedom in the diversity of all members of society. It is therefore not a member of society itself, but an instrument of production.

One could object that also humans could be mistreated as "animate things" like slaves, but this mistreatment is only possible directing purposes of labor or production – that is, in contexts beyond freedom. In slave societies, where humans are treated as "speaking instruments … the 'curse' of necessity remained a vivid reality" (Arendt 1998, 121). Conversely, if freedom is given, human plurality is entailed. Thus, dehumanizing people is only possible by ignoring their original integration in human plurality.

The fact that a language model like ChatGPT cannot agree on an opinion is demonstrated by the balanced indecision of its texts: If you ask ChatGPT to take a position on a topic, it will summarize opposing positions and draw a résumé of one or the other. Its texts are intended to be connectable, namely for different readers with opposing opinions, but it does not endorse any opinion itself. A distinction must be made between the fact that the language model can only develop its texts from the data available to it, and the fact that there is a moment when it seems as if it agrees with the opinions of others. Humans, too, can only form opinions based on their prior understanding and information, but their prior understanding assumptions do not determine who they agree with. Even if an AI-text generator prefers one option sharply like the model Delphi, this decision is not a result of negotiation but of sheer calculation.

Generative AI is now capable of something that is at the heart of Arendt's action: it can "speak." Although the main area of application is the creation of written texts, AI can now also speak by cloned voices. Is it a coincidence that Hannah Arendt hardly ever commented on writing, since for her, speaking, like acting, constitutes individual self-discovery in a social space of plurality? Arendt, on the other hand, seems to be as ambivalent about writing as Plato (Arendt 1978, 115), for whom writing does not make it possible to remember past thoughts but, conversely, to forget current thinking (Plato 2003, 274e – 275d). It almost seems as if Plato saw the "death of the author" (Barthes 1977, 142 – 148) coming two and a half thousand years ago. On the one hand, he wrote under pseudonyms and, on the other, he seemed to see the difference between the spoken and the written in the loss of the individual: When I speak, I am currently distinguishing myself from others; when I write, however, I am transferring my thoughts to a storage medium, so that it may even become unrecognizable who originally thought these thoughts.

The meaning of speech changes as soon as an AI speaks: it speaks in a human voice without the human in whose voice it speaks. Phenomenologically, this makes it unclear what the meaning of speech is and what it is. It appears through artificially intelligent voices *as* epoché – not simply *in the* epoché (see Husserl's explanation of epoché Husserl 1970, 77 (Hua VI, 78f)), because we can still mistake an artificially generated voice for a real one and take as true what it says. But we can no longer judge the meaning of the speech. We know then that we can be fundamentally confused about whether a human or an AI is speaking to us. When we speak, we can also take back what we say by claiming that not we were the ones, who said it. The meaning

of speech is then reduced. The spoken word can still be meaningful, but it is not possible to clearly specify what speaking means. On the one hand, we can in principle understand the difference between human speech and artificially intelligent speech. But since we cannot determine this difference with certainty in individual cases, it remains purely nominal. Whether it actually occurs in reality must then be left open (epoché).

We can experience and describe the accompanying phenomena of speaking, but we only experience them because we have to abstain from what speaking means here. "It" speaks, but what this speaking is, when neither the person we hear speaks nor what we hear is meant by the speaking instance, must remain open, because this speaking appears as this omission. It may not be what we think it is, and thus speaking is fundamentally no longer what it once was. For what speaking once was, speaking now stands as an omission.

If my claim is correct, an AI cannot act by speaking. If it is unclear who is speaking to me or what speaking means, then the processes that are part of acting do not take place: There is no free recognition between communication partners, there is no acting self-discrimination, because it remains open who the "self" is that is actually speaking to me.

But let's assume that we are also different when we write, because we give meaning to the texts: When we reproduce the same thoughts in a different situation, we follow roughly the same structure of thought. As a result, we identify *ourselves as individuals* with that direction of meaning. A language model lacks this ability to differentiate itself when writing: its function is to create texts that are as *human-like* as possible – and therefore not to differentiate itself from others. The text generator can only develop an individual profile if it is used regularly with one person, because the probability of syllable sequences also takes into account the internal communication with a user. However, if several people access the same account or use the same program on different computers, different profiles will result. It is not the text generator that is *self-differentiating* here, but rather it distinguishes different virtual profiles from each other, while not differentiating itself.

The virtual profiles, on the other hand, are external constructions: They do not have a self that could be different. Rather, their differences are the result of interaction. For this reason, the individual profiles of the text generation are just as incapable of action. They do not begin as something new and enter the world as individuals, but adapt to the users. Their individual profile is only a consequence of their use, a development along a chain of

causality, instead of starting something new, simply because it is a "solipsistic" tool which contradicts to Arendt's term of novelty which is embedded in social plurality. AI text generators are therefore not capable of acting in this way (speaking) or that way (writing) and do not create virtual beings capable of acting.

It could be argued that also people adapt to their social environment through regular interaction. The crucial point, however, is that these developments of self-differentiation from others are only possible because the child distinguishes itself from others from the very beginning. It individualizes itself only because it is already an individual. Its "process" of adaptation is individual, rather than becoming an individual through adaptation. It is not first a blank that becomes itself through social adaptation, for then it would never become itself because the social environment could not give it the space to be itself: It would remain trapped in the determinations of the social environment. This is not a psychological statement, but a logical implication of personal individuality that it cannot be an effect of a causal process. Thus, as soon as the social environment grants the individual selfhood space, it already perceives its self-distinction, which forces it to recognize it.

In contrast, an individual profile of a language model is a consequence of its adaptation to the human-machine interaction process. The human brings her individuality to this process. In addition, there is another reason why a language model is not capable of self-differentiation and therefore cannot act: The same language model is available to several users at the same time and develops different virtual profiles depending on its use. What apparent "self-differentiation" does it perform? In short, it is not different from others, but from itself. It is not identical with itself, because it develops different virtual profiles of text generation, none of which can be identified with itself.

Let us assume that language models interact with each other. Does this make them capable of acting in a quasi-social machine world? Is the machine-machine interaction a joint action? One could argue that individual adaptation to the machine-machine interaction process could not occur if the language models were not individual "at the outset" like humans. After all, if all interaction partners are not independent as machines, neither side brings any potential for individualization to the interaction. How then can virtual profiles emerge? From this objection one could conclude that individuality is either the result of environmental influences or that machines, like humans, can only individualize because they are individuals from the start. In the first

case, action becomes a sub-case of production (the environment influences a raw material until it becomes its product); in the second case, the "miracle" of the beginning is localized in an AI because it can compose texts.

In both cases, an AI's ability to act is measured by its characteristics: For the texts that an AI produces can also be produced monologically if it is programmed to give its commands to itself. In this way, it does not differ from others, but from itself. The crucial difference between the text creation of language models and human speech is that human individuality is not limited to the characteristics that distinguish one person from another. Rather, it requires another descriptive category. Hannah Arendt sums up the limits of what can be described in this way:

The moment we want to say *who* somebody is, our very vocabulary leads us astray into saying *what* he is; we get entangled in a description of qualities he necessarily shares with others like him; we begin to describe a type or a 'character' in the old meaning of the word, with the result that his specific uniqueness escapes us (Arendt 1998, 181).

The category that shows a person's individuality is that it *happens to* her and to us. However, the *character of being subject to events* (*Widerfahrnischarakter*) cannot be summed up in general characteristics. The fact that this character is without properties, and therefore cannot be described in the category of objectivity, is evident from the fact that a person is different from others *from the beginning*, even before it is possible to determine in what way she is different – for such a determination would only be the result of the individualization process from the very beginning. This is why I said at the beginning of my paper that community and individuality are mutually presupposed from the beginning, because a beginning, which is the human being, persists in community, but community consists in being a beginning.

In a machine-machine interaction, on the other hand, virtual profiles are only individuals from a human perspective, i.e. for the beings that interpret this interaction in a similar way as if it was humanlike. For machine-machine interaction, it makes no difference whether the same language model speaks to itself or to another. Rather, the input-output processing method is the same. The differentiation between several interacting machines, on the other hand, is done by humans. The different virtual profiles can be described *exhaustively* in terms of properties at any given time, even if these properties change during the course of the interaction. In this case, the properties determine the virtual profile. There is no quasi-social space for "joint" action between several machines, because there is neither a quasi-social space for new things,

nor do individuals interact with each other for whom a community exists. The technological unit takes the place of the social community.

IV. Should AI Act?

It becomes dangerous when the concept of action loses its contours to the extent that machines and computer programs are considered capable of action. This is where humans endanger their own scope for action. This happens when technological surveillance systems are built into human spheres of action, and when action is replaced by machine-machine interaction. Let's assume that democratic elections are abolished because artificial intelligence "knows" the political interests of citizens better than they do themselves, or makes accurate predictions about the political will of citizens for the coming legislative period. The political space would then be replaced by the category of production. This procedure can only be legitimized in a circular way by the AI certifying its own accuracy – albeit with the inductive proof that the predictions have always been correct in the past. But precisely because the argument here is based on inductive proof, which works in technical contexts, it underlines the fact that the political sphere of action is closed and replaced by production.

Even if a political mood can be predicted with certainty, this overrides freedom because, in case of doubt, the prediction must be enforced against the current will of the citizens. In this scenario, the accurate prediction is meant to ensure that the future political will of the majority is better estimated than the citizens themselves can imagine. Thus, if the correct prediction trumps the current political mood, interpersonal freedom must be suspended in the process (Harari 2017, 456).

Analogous dangers arise when AI speaks law, or when a techno-bureaucracy processes citizens' concerns with support of language models (Hermonies 2024, 336). In particular, when citizens' appeals against suspected erroneous rejections of applications are processed with ChatGPT just like the rejections themselves, the bureaucracy functions as a closed system of secrecy (Arendt 1979, 186) that leaves no room for joint communication – and thus no room for joint action. The danger of a loss of *interpersonal* action thus looms on all three levels of the state's separation of powers.

The "watchdog function" of political journalism is also at risk if it is replaced by AI-generated texts. Furthermore, the question arises as to whether journalistic reporting can still perform a political function of opinion-forming if the political will of citizens is no longer polled because their

interests can be more reliably researched by technology than through a referendum. The purpose of political reporting can then be reduced to journalism for the sake of pandering, which primarily offers entertainment value to citizens. Of course, one could argue that an AI can only determine the political will of citizens if they are also politically educated, i.e. have a political will. However, the crucial point in this scenario is that it is not the citizens who decide what their political will is. It is therefore within the technical control of the AI to what extent the political will of the citizens should depend on their political education.

In order to save the human sphere of action, free people must grant each other their political self-determination without it being occupied by AI predictions. No one should expect support for their own decisions to be optimized by an AI. Of course, we don't know if we'll still think our current referendums are right in a few months. But political freedom does not lie in our own inclinations, which change over time, but in the fact that we grant it to each other. Politics is not simply finding "the right answer" or "most efficient solution" to a problem; politics is what we do together and whose outcome cannot be foreseen and in which we are not determined in advance in either what we care about or how we want to achieve it together. The political space is where freedom shares unforeseen novelty and vice versa.

This also means that we accept to make a decision today that creates a binding situation that may outlast our own inclinations. Part of mutual freedom is that the freedom of political mandate holders is legitimized beyond our inclinations to shape the political space, and that this creative mandate is questioned by us at regular constitutional intervals. If an AI intervenes in this reciprocity, both the freedom of the elected representatives and the freedom of the voters collapse. If the AI knows today that in eight months, I will no longer agree with certain elected representatives that I would still vote for today, and therefore doesn't let me vote today, my political freedom is destroyed. At the same time, the mandate of a government depends on calculated probabilities and can be revoked at any time by recalculations. Such a government does not create political freedom, but implements an instruction manual. It does not act, it produces. Reciprocal political freedom therefore means giving each other temporary political roles, even if, in a few months, we are going to wish to have another government. The risk of being dissatisfied with the exercise of political power must therefore be borne in political freedom, because political freedom has nothing to do with optimizing production.

V. Result

AI cannot act and therefore should not occupy the human sphere of action. It should not give the impression that it can act – especially since even language models cannot write or speak in the sense of self-differentiation. People should cultivate their shared space for action by sparingly using interaction techniques where they do not know whether there is a person on the other side of the line. This includes a reluctance to engage in collective political decision-making in virtual social networks and a fundamental skepticism about whether the voice they hear online is really that of a human being. The process of mutual granting of freedom is most secure face to face. All digital medializations of discourse must be measured against this original situation of freedom. They must show to what extent they participate in this original situation from which they have digitally distanced themselves.

This is not an argument against receiving information through digital news channels or even taking note of written texts. But these activities do not automatically lead to action, because the mutual granting of freedom does not come into play. In the digital age, reading also belongs to the category of production – people "consume" information and optimize their own knowledge management. Not surprisingly, digital storage media are also used in this form of self-optimization. The only difference is that here people act monologically, without being dependent on mutual recognition of freedom.

Bibliography

ARENDT, H. (1977): *On Revolution*. London: Penguin.

ARENDT, H. (1978): *The Life of the Mind: The Groundbreaking Investigation on How We Think*. San Diego: Harcourt.

ARENDT, H. (1979): *The Origins of Totalitarianism*. San Diego: Harcourt Brace Jovanovich.

ARENDT, H. (1992): *Lectures of Kant's Political Philosophy*. Chicago: University of Chicago Press.

ARENDT, H. (1998): *The Human Condition*, Chicago: University of Chicago Press.

BARTHES, R. (1977): The Death of the Author. In: Barthes, R.: *Image Music Text*. Ed. by Stephen Heath. London: Fontana Press, 142 – 148.

HABERMAS, J. (1994): Hannah Arendt's Communications Concept of Power. In: Hinchman, L. P. – Hinchman, S. K. (Eds.): *Hannah Arendt*: *Critical Essays*. Albany, New York: SUNY, 211 – 230.

HARARI, N. Y. (2017): *Homo Deus: A Brief History of Tomorrow*. London: Vintage.

HERMONIES, F. (2024): KI in der Rechtswissenschaft: ChatGPT ernst nehmen? In: Schreiber, G. – Ohly, L. (Eds.): *KI:Text. Diskurse über KI-Textgeneratoren*. Berlin and Boston: De Gruyter, 329 – 340.

HUSSERL, E. (1970): *The Crisis of European Sciences and Transcendental Phenomenology*. Ed. by David Carr. Evanston, Illinois: Northwestern University Press (= Hua VI: *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie. Eine Einleitung in die phänomenologische Philosophie*. Ed. by Biemel, v. W. Den Haag 1954.)
PITKIN, H. F. (1994): Justice: On Relating Private and Public, In: Hinchman, L. P. –Hinchman, S. K. (Eds.): *Hannah Arendt: Critical Essays*. Albany, New York: SUNY, 261 – 288.
PLATO (2003): *Phaedrus.* Oxford: Oxford University Press.

Lukas Ohly
Fachbereich Evangelische Theologie
Goethe-Universität Frankfurt am Main
Norbert-Wollheim-Platz 1
D-60629 Frankfurt
Germany
e-mail: L.Ohly@em.uni-frankfurt.de