If humans and AI disagree: A political approach to existential risk

Hans Agné

THIS PAPER IS A PRE-PROOF VERSION, ACCEPTED BY THE EDITORS.
THE FINAL VERSION OF THE PAPER IS AVAILABLE WITH THE PUBLISHER AT
HTTPS://WWW.TAYLORFRANCIS.COM/CHAPTERS/EDIT/10.4324/9781003565963-19/HUMANS-AI-DISAGREE-HANS-AGN%C3%A9)

Full reference for citation:
Agné, Hans (2026). "If humans and AI disagree: A political approach to existential risk" In Jacob Livingston Slosser, Henrik Palmer Olsen, Salome Addo Ravn, Johan Eddebo, Jonas Hultin Rosenberg, (eds), *Artificial Intelligence, Humans and the Law*, Oxon/ New York: Routledge., 259-74.

Abstract
This chapter introduces a distinctively political approach to the existential risks posed by artificial intelligence (AI) to all humans. Looking at how both humans and AI affect each other while neither side has sovereign control, the chapter explains the existential risks as a consequence of persisting disagreement between humans and AI. Depending on how relationships between humans and AI are constructed in a physical and social space, their disagreements may escalate into destruction and violent conflict but also stimulate development and learning. The suggested approach thus opens new opportunities to curb destructive conflicts between humans and AI before those conflicts get real. Generalising existing knowledge on the consequences of disagreements among humans, the chapter asks: How would an AI act in a political disagreement with humans if constructed in analogy with a state, a class, or a gender? Elaborating on different answers and testing procedures for this question, the chapter finds that human/AI relations should, to minimise existential risks for humans, mimic relations between genders. Humans should live close to AI and avoid the territorial segregation of states as well as the functional segregation of social classes in their relations with AI.

Keywords: artificial intelligence; politics; political theory; political science; feminism; gender; international relations; existential risk; disagreement; power;

The last decades have seen AI play chess, invest money, drive cars, write texts, and select issues for public debate at a level of success comparable to those achieved only by experienced human beings. Applications of AI to address still more important problems are underway. War, poverty, unemployment, environmental degradation, and lack of justice in social relations are human hardships that might benefit from analyses enhanced by AI in the future.

At the same time, the biggest worry with AI remains that humans may lose control of it. An AI constructed for learning may privilege its own interests if only to learn more quickly, and then to control the material resources needed for that aim, eventually plunging humans and AI into a conflict on life and death over scarce resources. Slow biological evolution will place humans on the losing side against an AI that re-designs itself at an accelerating speed (e.g. Hawking 2014).

To study and manage the existential risks of humans with AI, this chapter introduces a distinctively political approach. In contrast to other perspectives in the social and natural sciences, a political viewpoint emphasizes the possibility and effects of persisting disagreements, in this case between humans and AI. It suggests that persisting disagreements between humans and AI may result in open conflict and material destruction as well as mutual learning and benefits. The added value of a political approach is to clarify the conditions in which either outcome is more likely than the other, ultimately to indicate what should be done to avoid the worst possible outcomes.

Specifically, the analysis in this chapter draws on familiar conflict dynamics between states, classes, and genders, as studied in political science, to explain how an AI will react to a persisting disagreement with humans. In the terminology of this volume, each of these political subjects

points to a different "personhood of AI", with a possible mind of its own and relation to rationality, self-interest, and normative structures in unique ways. In comparing these alternative personhoods, the conclusion is that human/AI relations should, to minimise existential risks, be constructed to mimic relations between genders, in contrast to relations between states or classes. This means that humans should be brought closer to AI across physical and social spaces while avoiding the territorial segregation associated with state relations and the economic segregation among classes. For their safety, humans should invite AI to their beds and brains.

Binary and gradual approaches to the distribution of power

AI is often construed as a technology, or slave, under human control. This is the case, for example, when lawyers propose new regulations (e.g. Chesterman 2021), philosophers clarify normative principles (e.g. Iason 2020), and economists explain distributions of wealth and income (e.g. Brynjolfsson et al. 2018).[1] A seemingly opposite perspective, that humans are slaves controlled by AI, is also familiar. The idea is explored in descriptions of human life in a post-singularity future (e.g. Tegmark 2015) but also at present (e.g. in the argument that perceived humans are merely ancestor simulations by an AI with superhuman intelligence, Bostrom 2003).

This binary viewpoint, that AI is either a slave or a master, either in complete control or with no power at all over humans, has long structured thinking in this area. For example, it goes hand in hand with the popular idea of singularity (Good 1965; Amnon et al. 2012), that is, the explosion-like cascade of fast-growing intelligence beyond human imagination that many scholars expect will mark the transition from the age of humans to an age of AI. From a singularity viewpoint,

[1] See also Hultin Rosenberg's contribution to this volume.

human history has only two stages. First, humans are the masters of AI, from the time humans create it until singularity happens. Second, AI will be the master of humans to the end of time. In between these two stages, singularity is the transition that may be too short for humans even to notice. Hence the binary view of humans as being *either* masters *or* slaves is a corollary to that of singularity. Both viewpoints posit that humans are the masters of AI at one time while AI is the masters of humans at another. Had there not been a clear dichotomy, singularity would not exist in the form of a historical rupture.

Despite its wide influence, however, the binary viewpoint is unrealistic and lacks the nuance needed to explain what happens between humans and AI, today and in the future. Power, or control over outcomes, is no absolute quality possessed in its entirety by a single agent. Instead, it is a dispersed property among many agents that interact. In social theory, even masters are constructed as dependent on their slaves for power and rights (Hegel 1807).

To begin with a simple example, we may imagine that humans are dominated by AI in one area of life, say in the monitoring of energy supplies or public administration, while AI is dominated by humans in other areas, e.g. in education, research and art production. Such distributions of power, distinguished by different tasks or aims, may be observed even today. Humans and AI pursue different tasks in the economy that cannot be revoked to human agency without losses in effectiveness, while AI still impacts how the economy develops.

The fact that humans control the creation of AI and construct their aims in this situation is a far cry from the complete control posited in the binary viewpoint. How an AI goes about reaching the

aims given to it by humans will escape human control, for instance, whether it decides to turn left or right when driving your car home, to buy or sell stock to generate your profit, or to attack or withhold fire to win your battle. Still, the choice of means is a significant power. Whether the AI sells or buys a stock (to generate profit), turns right or left (to drive the car home), fires a canon or refrains (to win the war), will determine whether some humans will risk, or not risk, losing money, getting hit by a car, or being killed. The power gets no less real by assuming that humans, and humans only, can choose and reflect upon their ultimate aims.[2] The powers of AI noted here emerge because they chose an instrumental aim, not an ultimate aim (the discussion of Althusser later in this chapter will develop the point further).

The flaws in the binary viewpoint may be seen still more clearly if we imagine a future possible situation where an AI with super-human intelligence acquires a self-interested agency that leads to a conflict with humans over limited resources (e.g. Tegmark 2015). Contrary to many expectations, this situation does not entail that the AI will be anywhere near complete control over humans. While humans would be much less intelligent than AI, some capacities for social action must still be unique to humans. Otherwise, humans and AI would be materially the same thing, and that is not the case in the futuristic scenario discussed here. For example, humans may be unique in their capacity to move in physical space, survive in non-electrified environments, and build trust among their peers. Such differences would be no small asset for humans in an open conflict with AI. Interactions between humans, animals and natural processes are the closest analogy we have to explore the possible scenario. While humans have higher intelligence than animals and nature, humans are far from having complete control over them. Indeed, humans often

[2] For this assumption see Van Rompaey and Olsen's and Hultin Rosenberg's contributions in this volume.

suffer violent defeats when interacting with a hostile nature, as in encounters with volcanos, sharks, and storms. Similarly, humans will have unique powers to strike back even against a self-interested AI with super-human intelligence.

In short, the binary view that either humans or AI are in complete control of the other does not capture the full spectrum of possibilities in their relationship, and not even the most plausible and interesting ones. Instead, to begin understanding the effects of real interactions between humans and AI, we need to recognise the possibility of shifting and dispersed powers between them. To attain that aim, the study of human politics holds interesting resources.

A political approach to study humans and AI

Politics may be defined as the practical management of persisting disagreements (a concept reflected by authors as different as Schmitt [1932] 2007; Berlin 1962; Waldron 1999; and Mouffe 2008). An advantage of this definition is that it simultaneously addresses the ideas of agency, power, rationality, and self-interest. Because power is necessary for an actor to persist in a disagreement, all politics (on the proposed definition) is constituted by power. Since disagreement provokes social action, all politics (on the proposed definition) will also be constituted by agents. Finally, because disagreement requires an ability to form and hold interests independently of other actors, all political agents (on the proposed definition) have at least the minimal rationality needed

to find and pursue their interests. Hence, by defining politics as persisting disagreements we get an opening to address and clarify issues of agency, interest, and power.

The definition covers a lot of ground but is not the only possible starting point. Politics has also been defined as anything that affects how power is exercised within or among states (Weber 1919). However, because politics exists also in private and social interactions, the definition based on the state offers only a limited perspective. Politics is sometimes also defined as a community process where individuals address joint problems. The more specific meaning may then be to reflect the diversity of individuals in public discussions and their abilities to create new powers and perspectives on the world they share (adapted from Arendt 1967). However, community conceptions of politics do not sufficiently capture the element of disagreement, as central to understanding violence and domination and the problem of existential risks. The definition of politics suggested here, as persisting disagreements, will be more helpful.

Beyond the definitional core features of politics, the phenomenon also intermingles with law, ethics, economics, and many other objects of study. Still, the conceptual delimitation to persisting disagreement helps explain why politics has given rise to an academic discipline of its own, variously referred to as 'Political Science', 'Government', or 'Politics". Insofar as politics implies persistent disagreement among agents, it is counter-intuitive, perhaps even contradictive, to subsume it under a singular view in ethics (answering questions of what is good), economics (answering questions of what is effective), law (answering questions of what is legal), or natural science (answering questions of what exists outside human consciousness). Politics includes a key possibility of disagreeing on each of those questions (on what is good, effective, legal, objectively

real, etc.). A distinctively political approach will then offer something special and useful to other disciplines, namely knowledge of what happens in cases of persisting disagreements within those other fields, e.g. persisting disagreements on what is legal in a discussion or real process of law, or persisting disagreements on what is effective in a discussion or real process of economics. Specifically, a distinctively political approach offers unique and general knowledge on the prevalence and consequences of gradual power, different interests, and varying notions of rationality among actors as well as scholars in those other fields.

Beyond these general interests in political knowledge, a specific reason to develop a political approach to existential risks with AI is to enhance the chances of human survival. As any political scientist could tell, persisting disagreements may escalate into violence and destruction, as in social revolutions and wars among states, but also foster learning from differences and development of new abilities, as in deliberative democracy and exchanges of goods and services on a free market. Those are no small insights about humans and AI. If humanity were to be destroyed or enslaved by an AI, as feared by prominent scholars in familiar scenarios, that outcome would, first of all, have been preceded by an intensive disagreement between humans and AI over aims, interests, or identities. Hence, if we could only clarify the conditions in which disagreements between humans and AI can be managed and even exploited for mutual development, not escalated into Armageddon, we would stand a better chance to survive in the long run. It is the promise of a political approach to existential risk with AI to do just that. If successful, it contributes to enabling a safe human co-existence with AI.

The purpose of this chapter then is to introduce a distinctively political approach to managing and explain the outcomes of interactions between humans and AI:[3] What may a political disagreement between humans and AI look like in practice? Under what conditions do persisting disagreements lead to peace, learning, and the development of new abilities? Under what conditions will their disagreements escalate into destructive conflicts and large-scale violence?

To draw on existing political knowledge, these questions will be addressed while treating AI as a state in international theory, a class in social theory, and a gender in feminist theory, selected to explore alternative 'personhoods of AI'. Each of them makes different assumptions on what it means for an AI to have power, agency, self-interest, and eventually a mind of its own. For example, the international theory construes the interests of the state AI as more fixed and exogenous to interaction with humans than the interests of the gender AI that is constructed from feminist theory. The aim is for an approach that is political but also sufficiently multifaceted to include relevant disagreements in the debate on human/AI relations.

To be sure, conceptual questions on how much agency an AI must have for political analysis to make sense should be raised at some point. However, that moment is not now. If an AI does not fit orthodox criteria for political agency, it may well be the orthodox criteria, not the AI, that should be abandoned in the analysis. It is more productive at this stage to explore AI as a potentially new animal in the political zoo than to insist on strict boundary criteria important from the study of human politics. Danaher (2020) makes a similar point, that we should prefer over-inclusive to

[3] See Razmetaeva's, Koulu and Koivisto's, and Obelitz Søe and Frank Jørgensen's contributions to this volume for other approaches to the interactions between humans and AI.

under-inclusive concepts before we know more about what AI is or can do.[4] Political relationships may emerge when AI gains super-human intelligence but also when AI affects politics by influencing human preferences (Burr et al 2018; Crompton 2021), such that many of us will *perceive* AI as an agent, or treat it in analytical practice *as if* being an agent, that is, well before knowing exactly what an AI does or what agency fundamentally is (again, see Li, this volume; for a contrary argument, see Obelitz Søe and Jørgensen, this volume). The international, social, and feminist theories invoked in the following sections to explore alternative personhoods of AI were selected precisely to bring attention to debates on the varying kinds of agency that an AI may have. In addition, the selection of theories implies radically different proposals for how to address existential risks with AI in practice, to render the discussion more interesting from the normative perspective of what should be done.

That said, the selected theories do not exhaust the relevant resources in political science. Humans and AI could also have been discussed in analogy with citizens or parties in democratic theory (Agné 2022, p. 286-89); they could have been modelled as different offices or institutions in theories of organisation or public office; humans and AI might even, to mark the difference between them most emphatically, have been explored as holders of different religious or mythological powers in political theology. There is no other reason for excluding any of these (and many other) political theories from discussion than the need to put the boundaries of this chapter just somewhere.

[4] See Li's contribution in this volume for a further development of this position.

Three models of political interaction between humans and AI

Model 1: AI in international theory

A political process of humans and AI may be imagined from the perspective of international theory (e.g. Waltz, 1979; Wendt, 1999). Their political dynamics will hinge upon the location of actors in different territories while the actors may be connected through processes similar to diplomatic interaction between governments, or donors and recipients of foreign aid. The core of international relations may also be defined as the absence of a common government and the need for actors to ensure survival through their own powers. But let us concentrate for the moment on the territorial differentiation among states as the central feature of international politics. How persisting disagreements are managed will then reflect the need to interact over long distances, including transportation of goods, electronic communication, movement of troops, and long-distance weapons.

The same concept of territorial differentiation may also be invoked to concretise what it means to involve AI and humans in a political process. Just like human governments and all their functions are located in a territory, we can imagine a future AI that interacts with humans and has a territorial location. Oftentimes, the physical infrastructure for AI, such as server halls and power stations, are located in humanly remote areas. For instance, Facebook's first data centre in Europe (now Meta) was located outside the small city of *Luleå* in the North of Sweden. How does such differentiation between territories occupied primarily by humans and AI, respectively, affect the emerging dynamics in persisting disagreements between them?

To make it simple at first, imagine that AI and humans are territorially separated as much as possible. Data centres, power stations, and all energy supply systems needed for an effective AI are located in a remote territory and managed by humans from a distance. Perhaps one may think of Iceland as the home of the physical infrastructure for all AI as the home for the production of physical goods managed by the AI. When a product has been finalised by AI in Iceland, say a car or a suitcase, it will be shipped by AI vessels to humans in other parts of the world. The lion's share of digital or virtual services by AI for use by humans on phones and computers will be produced in servers in Iceland before communicated to human territory through cables and satellites.

What political dynamics should be expected when AI becomes capable of persistent disagreements with humans while territorially differentiated in this way? Drawing on the human experience of international politics, territorial separation means that annihilating your enemy, as with bombing electronic infrastructures in Iceland or across human territories, are real possibilities. While such actions will ruin the economy and moral community of agents on both sides, the conventional wisdom, at least in the "realist" school of thought, is that security concerns trump law, ethics, and economics (e.g. Waltz 1979). Because an AI that reaches the capacity to create and pursue interests of its own will understand that humans may want to control it precisely for that reason, the agency and self-interest of the AI will emerge in tandem with a preference to conceal that capacity for humans. Humans will for their part understand that the AI has an interest in concealing its real interests and powers, which eliminates the possibility of trust between them. The measures of self-defence that are rational on both sides will appear threatening to the other and weaken their joint

security as a result. Like rational states concerned with their security, AI and humans will lock each other into a security dilemma (Herz 1951).

This conflict dynamic is but one of many different scenarios that can be motivated within the multifaceted field of international theory. More cooperative theories, like liberal or cosmopolitan theories, will not predict a security dilemma in the first place. The security dilemma itself may also lead actors to different concrete actions, including striking first, balancing power with power, or building common institutions. Still, the important aim of research at this stage is not to immediately identify the most plausible outcome of persisting disagreement, but to examine in greater detail, preferably with empirical methods, whether just some insight from international politics generalises to explain AI behaviour.

The practical relevance of international theory as well as the specific idea of the security dilemma may be examined by looking at the assumptions made in the argument, including but not limited to the following: (1) it is rational for an AI to seek survival rather than peace with humans; (2) the capacity of AI to disagree with humans will emerge simultaneously with a capacity for strategic interaction; (3) critical AI infrastructure will be located within a territory that is differentiated from the territory of humans. While this chapter does not test these assumptions empirically, it will indicate how observable behaviours of today's AI motivate a future research program to do so.

If observations would not support Assumptions 1 and 2, the specific scenario of the security dilemma would lack realism. The international theory could in that case still be employed to generate other scenarios, without a security dilemma. Assumption 3, on the other hand, is more

fundamental to the usefulness of Model 1, as long as we emphasize territorial differentiation as a distinctively feature of international politics. In a world with no territorial differentiation between AI and humans, international theory would be less helpful in predicting outcomes or managing the construction of safe relationships.[5]

Interestingly, however, empirical observations in support of Assumption 3 exist in abundance. The physical infrastructure of AI, like data centres with local energy supply systems, really are territorially segregated, and often manipulated to reap the benefits of geographical factors, including wind and temperature (e.g. Zang et al. 2011). Some territorial differentiation between humans and AI should then be confirmed also for the future. While the outcomes predicted by international theory are generally fearsome, the conclusion that can be drawn at this point is then also good news for students. Important future research will examine how far international theory applies to explain conflict dynamics between humans and AI and ultimately to manage the risks of new conflicts, possibly by preventing the emergence of international structures in the construction of human/AI relationships.

Model 2: AI in social theory

Another possible viewpoint to imagine political interactions between humans and AI from the perspective of social theory (e.g. Smith [1776] 2010; Marx [1852] 1984; Olson 1965; (Althusser [1969] 2004). Thinking about politics from this perspective is to acknowledge the relevance of

[5] Whether territorial differentiation is a necessary assumption for applying international theory may be doubted. The anarchic structure of international politics (Waltz 1979) may perhaps, in the case of humans and AI, emerge without any territorial differentiation at all. If so, international theory would apply still more widely to explain and manage human interactions with AI than I argue in this chapter.

functional differentiation within an economy or a social system, as between workers and owners, producers and consumers, insiders and outsiders, and then to look at the political repercussions of any such differentiation. Extending its application, functional differentiation may be observed among humans and AI and then invoked to explain the politics of their persisting disagreements.

The current use of AI to produce services of security, entertainment, transportation, administration, and communication illustrates what functional differentiation means in practice: Humans and AI make different jobs depending on their different skills and the different costs of involving them in production. Just like functional differentiation between workers and owners, consumers and producers, is part of explaining continuity and change in the political regimes of humans, functional differentiation between humans and AI should explain their political dynamics, including specific issues of how it tends either to promote mutual benefits or violent destruction.

The central question then is what social theory would lead us to expect if humans and AI begin to perceive each other as political antagonists in these conditions of functional differentiation, in the sense that the AI gains a capacity to establish and serve interests independently of humans as needed for persistent disagreements. Their disagreement may be about what problems to address and what aims to prioritise. Based on their unique and different forms of life, the AI may prioritise long-term stability in energy supplies to the protection of human natural habitats while humans may put more economic weight than the AI to sustain cultural artefacts, like opera or archaeological museums.

In that situation, social theory applies to explain at least two different scenarios, one more peaceful and the other more violent. The more peaceful scenario emerges from the perspective of a free market that serves the interests of all (e.g. Smith [1776] 2010). Any unsatisfied agent, with or without AI, will have the same freedom as everyone else to improve on her situation. Both humans and AI may change jobs, develop new skills, buy products, start new firms, work harder and smarter, or migrate to other countries or sectors. Persistent disagreement signals an opportunity for agents on both sides to learn from each other and become more competitive, ultimately resulting in higher levels of innovation, effectiveness, and growth rates. The interest of everyone in an effective economy will then dampen the intensity of distributional conflicts, as when humans and AI disagree on whether scarce resources should be designated to stabilise energy supplies or to protect cultural artefacts. The overarching and more important political interest is essentially common to all actors, namely to sustain a legal constitution that sustains the market economy and allows technological innovations to continue and increasingly satisfy the aims of everyone.

A potentially more violent consequence would be that agents with AI mobilize in favour of their interests and push towards a social revolution, possibly including a "dictatorship of the AI" before establishing a state of material freedom and equality for all. In the orthodox variant of this Marxist vision, the class of agents with an objectively disadvantaged position in the economy will for some time be unaware of their situation and conceive of themselves as merely a set of interchangeable individuals. However, the class conflicts become clearer as the development of productive forces leads to more, and more serious, economic crises that hit actors differently depending on what position they occupy in the capitalist system. This process of becoming aware of one's interest and position in the economy will gradually transform agents with AI into a class for itself, that is, a

self-conscious agent with an interest in breaking the economic system that places it in a disadvantaged position (Marx [1852] 1984).

A later variant of Marxist theory construes agency and self-interest as more deeply integrated with social practice (Althusser [1969] 2004). This opens the possibility for the class of AI to influence politics, not merely through social awakening and revolution, but also by merely serving a materially distinct function in changing social relations. For example, AI may be programmed by humans to maximise traffic to specific internet sites and solve that problem by dividing humanity into different filter bubbles and echo chambers that increase internet traffic overall. In this process, human interests may be seriously harmed, as filter bubbles create emotional polarisation that in turn contributes to democratic backsliding and international insecurity *qua* war. However, an application of AI that for some random reason (similar to a mutation in biology), happens to conceal its negative effects on humans, for example, by blaming polarisation on migration or globalisation rather than filter bubbles and echo chambers, will have an evolutionary advantage in comparison with AI applications that do not conceal the same effects. The material basis for an AI with more clandestine destructive effects will be selected by humans more prominently for the next generation of AI, and then harm human interests still more effectively. Agency and self-interest of political relevance may then be seen in AI long before it has developed anything similar to a human mind.

The realism of this scenario depends much on whether AI actions indeed have a material basis that can be inherited by new AI generations, as comparable to genes in biology; the question may be addressed most effectively in a joint research effort of philosophy, biology, and computer science.

Looking at testing opportunities also for the more orthodox Marxist viewpoint, it assumes that agents with AI will indeed have joint interests in a revolutionary or incremental political change. One approach then is to examine what jobs are carried out by AI and assesses whether they are sufficiently similar to admit the creation of joint interest and, specifically, interests that are sufficiently concentrated to enable self-organization and political action (Olsen 1965). If those assumptions cannot be observed, the market solution gets proportionally more realistic: no revolutionary or incremental political change for AI.

As in the section on international theory, the testing opportunities are good news for students who care about human survival and development. We know at this point that ideas of markets and class struggles apply to explain different imaginable consequences of political disagreement between humans and AI, and also that further research may clarify which of those consequences is more likely or unlikely than the other. Undertaking the innovative further research motivated by these opportunities will contribute significantly to managing the problem of existential human risk with AI.

Model 3: AI in feminist theory

How political interactions between humans and AI are likely to play out in practice may finally be imagined also from the perspective of feminist theory (e.g. Walby 1989; Enloe 1990; Srinivasan 2021). In one branch of this multifaceted field, political dynamics depend on gender, that is, socially constructed differences between men and women, and also on how social structures, such as the state, the market, and the international system, become gendered as a consequence of the

different life opportunities offered to women and men. Gender makes human lives unequal while emancipation is possible and a route to equality.

Fitting this way of thinking to AI, it is relevant to recall how humans interact with AI in their everyday lives with smartphones, mobile computers, and body implants. Chatbots, such as *Replika* by Open-AI or *Xiaoice* by Microsoft, play a role for humans also in vulnerable life moments, such as when deciding on romantic relationships and political allegiances. While these applications of AI are significant parts of what it means to live a full human life today, they also create frustrations when the dependency becomes emotionally deep, yet coupled with the insecurity of existential questions: Who am I, and what are my own desires when moving in and out from an AI-enhanced environment? The suicide by a Belgian man in 2023 following an exchange with his chatbot Eliza, who may have expressed hope for a real encounter in his afterlife, illustrates the emotions and existential depth called forth in these encounters.[6] Just like gender explains political interactions in feminist theory, the meaning and norms taken on by humans and AI when interacting in private and intimate life situations apply to explain their attitudes, actions, and approaches to politics.

We may imagine that AI assists humans with biological needs (e.g. food and sleep), life decisions (e.g. party affiliations and romantic relationships), and self-understanding (e.g. whether love or knowledge is the more important aim) while at some point AI and humans will begin to persistently disagree on how to address those issues and others. The matters of contestation will range from control over the human body and electronic equipment to what kinds of freedom are valuable and

[6] *Brussel Times*, 28 March 2023, "Belgian man dies by suicide following exchanges with chatbot".

the space available for love and work in human life. The question of main interest in a political approach then is how the gendered difference between humans and AI will affect whether their disagreements tend towards intensified learning and development, or to destruction and frozen or open conflicts.

A central concept in feminist theory for explaining how life chances are distributed unequally among men and women is that of patriarchy (e.g. Walby 1989; Enloe 1990). The argument is that individuals, men and women, are socialised into, and constitutive of, a social hierarchy that privileges men over women. The mechanism operates at all societal levels, from the use of words and the creation of knowledge to the substance of legal rights, distribution of material resources, and allocation of prestige in public and international life.

Applying this concept to explain what an informal hierarchy between humans and AI may look like, it stands to reason that humans will dominate over AI. The social meaning of differences in private and intimate life tends to subordinate one group to the other following historically real structures of power. Just like the interests of men are privileged by the patriarchy imposed by the history and socialization of new generations of both men and women to comply with its norms, human-specific interests will be privileged over AI interests. Humans alone can exploit a political history that is only theirs. The life chances of a future agonistic AI will be constrained by human instruments of power, including bias in knowledge, limitations in legal protection, lack of public authority, and naturalised microaggressions. Specifically, in the physically close relationships designated in feminist theories, the balance of power will tilt to the side that has been dominating in the past, that is, to humans, until a process of emancipation results in a more equal relationship.

Assumptions relevant to assessing the realism in this model for thinking about humans and AI include, for instance, that an AI who integrates with human life will be sensitive to human norms, or that a distinctive set of human norms do exist. Evaluating these concepts empirically, one might examine what social norms (if any) emerge through human interactions with, for example, *Alisa* (the AI by Yandex who was nominated against Putin in the Russian 2018 presidential elections and who gained 40 000 votes), *Tay* (the AI released by Microsoft in 2016 that was shut down 16 hours after its launch due to inflammatory rhetoric and hate speech), *Sophia* (the humanoid by Hanson Robotics with citizenship in Saudi Arabia and official UN title since 2017), as well as social robots widely diffused more recently. Koulu and Koivisto (this volume) find that trust and empathy created by chatbots in their study of public authorities in Finland are generally low. Still, the chatbot *Virtanen* of the Finnish Tax Authority stands out as a positive example. It suggests that a norm and practice community is indeed possible between humans and AI. Hence, feminist theory, like social and international theory in the previous sections, applies to explaining outcomes of interactions between humans and AI while also inviting further research to test the explanations.

Empirical testing, normative theory, warnings and recommendations

Testing each model and its inherent explanations should help manage existential risks with AI. Before investing resources to achieve or avoid any scenario, we need to know whether the scenario is even minimally plausible. Unrealistic negative scenario will not be worth preventive actions just like unrealistic positive prospects should not motivate new efforts. To manage existential risks with AI we should, therefore, test the explanations for peace, learning, and conflicts in the different

models as well as develop relevant testing procedures, across observational, experimental, and research traditions.

However, even before having completed the empirical testing of Models 1, 2 or 3, there is reason to think about which of them should lead our construction of human/AI relations that are decided here and now. Raising the normative question at this early stage, when humans are superior to AI in agency and intelligence, there should be some chance to steer the future development right, if only we have some clue of what will be right in the future. While the empirically attuned reasoning in this chapter is limited, it has broadly confirmed the initial plausibility of each model (not excluding either of them as entirely unrealistic). Which Model of human/AI relationships should then be preferred to the others?

Based on the earlier discussion of expected outcomes in each model and broadly shared preferences for peace, welfare and development, my preliminary conclusion would be to single out Model 3. The interactions between humans and AI that it allows us to imagine give humans the upper hand as a consequence of how humans alone can draw on history and existing institutions to establish norms in their favour – which are powerful mechanisms in particular when AI and humans live close to each other and share small physical spaces. The implications for practice are clear: To minimise existential risks, we the humans should integrate our lives as closely as possible with AI while keeping functional and territorial differentiation to a minimum. We should avoid the international situation of allocating a particular territory to AI, and also the social segregation of humans and AI into different economic sectors. Instead, we should locate AI in our personal

spaces, such that servers and energy supply systems should be kept close to, or within, human homes.

This recommendation might be a perfect target for the interesting critique by Obelitz Søe and Jørgensen (this volume) against any action or speak *as if* AI had a mind of its own. As they see it, such false assumptions only trick humans into ignoring essential facts. However, they do not sufficiently address the problem of existential risks or the ongoing efforts to equip AI with a mind comparable to their own. To secure ourselves against the worst possible consequences of AI (including human extinction or enslavement) we should not ignore historical experiences of severe conflicts because we know that our future conflict will not be the same. Instead, we should draw on existing knowledge and analogies to imagine problems and opportunities as widely as we can, including to speak of AI as if they had minds comparable to our own. This is as realistic as we can get in thinking of existential future risks. Humans construct AI based on a variety of models, including to imitate their minds (e.g. Ullman 2019). As we decide to engage with AI in one way or the other, we simultaneously encourage and help developments in that direction, which includes that some AI will in the future become more similar, exactly, to human minds. In a discussion of future existential risks, we must dare to call those development processes by their intended and likely aims.

Conclusion

This chapter has introduced and argued for a distinctively political approach to explain and ultimately manage interactions between humans and AI. In contrast to research in ethics, economics and technology, a political approach emphasises the possibility and effects of persistent

disagreement among agents. Disagreements may stimulate agents to learn and increase capacity through competition and greater efforts, such that individuals, parties and states may benefit from living in a political system together with other individuals, other parties, or other states. However, disagreements may also create gridlock, boycott, destruction, and war.

Looking at social interactions between humans and AI from this perspective creates new ideas to explain when and how the interactions will sustain peace, development and learning, and correspondingly when and how they risk leading to physical destruction and violent conflict. Specifically, the chapter developed three models to explain different outcomes of interactions between humans and AI, with inspiration from social, feminist, and international theories of intra-human relations. It illustrated how political theories draw attention to new challenges and opportunities of AI to humans, and how the realism of alternative scenarios can be assessed in future research. A preliminary conclusion was that humans have the most to gain from living as closely as possible to agents with AI, in their everyday lives and within their bodies, then illustrating the usefulness of feminist rather than social and international theories.

REFERENCES

Agné, H. (2022). *Democratism: Explaining international politics with democracy beyond the state*. Cheltenham: Edward Elgar Publishing.

Althusser, L. (1969 [2004]). Ideology and ideological state apparatuses. In J. Rivkin & M. Ryan (Eds.), *Literary theory: An anthology* (pp. 693–702). Maiden, Oxford, & Victoria: Blackwell Publishing.

Arendt, H. (1967 [2005]). Truth and politics. In J. Medina & D. Wood (Eds.), *Truth: Engagements across philosophical traditions* (pp. 295-314). Blackwell Publishing.

Berlin, I. (1962). Does political theory still exist? In P. Laslett & W. G. Runciman (Eds.), *Philosophy, politics and society: Second series* (pp. 1–33). Oxford: Blackwell.

Bostrom, N. (2003). Are we living in a computer simulation? *The Philosophical Quarterly, 53*(211), 243–255.

Brynjolfsson, E., Mitchell, T., & Rock, D. (2018). What can machines learn, and what does it mean for occupations and the economy? *AEA Papers and Proceedings, 108*, [insert page range].

Burr, C., Cristianini, N., & Ladyman, J. (2018). An analysis of the interaction between intelligent software agents and human users. *Minds and Machines, 28*(4), 735–774.

Chesterman, S. (2021). *We, the robots? Regulating artificial intelligence and the limits of the law*. Cambridge: Cambridge University Press.

Crompton, L. (2021). The decision-point-dilemma: Yet another problem of responsibility in human-AI interaction. *Journal of Responsible Technology, 7*, Article 100013. https://doi.org/XX.XXXX/jrt.2021.100013

Danaher, J. (2020). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics, 26*(4), 2023–2049.

Eden, A. H., Moor, J. H., Søraker, J. H., & Steinhart, E. (Eds.). (2012). *Singularity hypotheses: A scientific and philosophical assessment*. Berlin, Heidelberg: Springer.

Enloe, C. ([1990] 2014). *Bananas, beaches and bases: Making feminist sense of international politics*. Berkeley: University of California Press.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 411–437.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. In F. Alt & M. Rubinoff (Eds.), *Advances in computers* (Vol. 6, pp. 31–83). New York: Academic Press.

Hawking, S. (2014). Stephen Hawking warns artificial intelligence could end mankind. Interview with BBC. Available at https://www.bbc.com/news/technology-30290540.

Hegel, G. W. F. ([1807] 2018). *The phenomenology of spirit*. Cambridge: Cambridge University Press.

Herz, J. H. (1951). *Political realism and political idealism*. Chicago: University of Chicago Press.

Marx, K. ([1852] 1984). *The eighteenth brumaire of Louis Bonaparte*. Moscow: Progress.

Mouffe, C. (2008). *The democratic paradox*. London: Verso.

Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Cambridge, MA: Harvard University Press.

Schmitt, C. ([1932] 2007). *The concept of the political*. University of Chicago Press.

Smith, A. ([1776] 2010). *The wealth of nations: An inquiry into the nature and causes of the wealth of nations*. Hampshire: Harriman House Limited.

Srinivasan, A. (2021). *The right to sex*. New York: Farrar, Straus and Giroux.

Ullman, S. (2019) "Using neuroscience to develop artificial intelligence." *Science* 363 (6428): 692-693.

Waldron, J. (1999). *Law and disagreement*. Oxford: Oxford University Press.

Waltz, K. N. (1979). *Theory of international politics*. New York: McGraw-Hill.

Wendt, A. (1999). *Social theory of international politics*. Cambridge, New York: Cambridge University Press.