ATTILA GYULAI, ANNA UJLAKI

The political AI: A realist account of AI regulation

This article adopts a political theoretical perspective to address the problem of AI regulation. By disregarding the political problem of enforceability, it is argued that the applied ethics approach dominant in the discussions on AI regulation is incomplete. Applying realist political theory, the article demonstrates how prescriptive accounts of the development, use, and functioning of AI are necessarily political. First, the political nature of the problem is investigated by focusing on the use of AI in politics on the one hand and the political nature of the AI regulation problem on the other. Second, the article claims that by revisiting some of the oldest political and theoretical questions, the discourse on guidelines and regulation can be enriched through the adoption of AGI and superintelligence as tools for political theoretical inquiry.

Keywords: *Artificial Intelligence, Political Theory, Political Realism, Applied Ethics, Enforceability*

Author Information

Attila Gyulai, Centre for Social Sciences - Institute for Political Science / University of Public Service
https://orcid.org/0000-0003-2471-6049

Anna Ujlaki, Centre for Social Sciences – Institute for Political Science / Corvinus University of Budapest
https://orcid.org/0000-0002-8030-624X

How to cite this article:
Gyulai, Attila, Anna Ujlaki. “The political AI: A realist account of AI regulation.”
Információs Társadalom XXI, no. 2 (2021): 29–42.
https://dx.doi.org/10.22503/inftars.XXI.2021.2.3

All materials
published in this journal are licenced
as CC-by-nc-nd 4.0

1. Introduction

With the widespread emergence of artificial intelligence in the modern world, decision-makers have become increasingly aware that the development, use and functioning of AI require some level of regulation. From AI companies and research institutions to governments and supranational organizations, several actors in the field recognized that despite the proliferation of self-imposed value-systems and guidelines (Hagendorff 2020; Héder 2020), and even if the value-alignment and value-loading problems (Christian 2020; Bostrom 2014) become solved, the question of how the desired values and norms could be realized remains open. Although some researchers think that the race for AI will result in a race for regulation (Smuha 2021), or that the ethical frameworks and the users can be brought closer to each other (Hatamleh and Tilesch 2020), thus allowing the guidance problem to be settled by taking the values and norms into account on the one hand and the developers and users on the other, we argue that a third factor, namely politics should be included in the discussion.

In this article, we argue that the problem of implementing AI guidelines, whether they concern narrow AI, expert systems or superintelligence, will be necessarily focused on the question of enforceability. That is, whereas the discussion on guidelines concerns mostly their value content and the norms they are aimed at applying, our claim is that the issue of how they can become binding should be taken into account more seriously. By its nature, it is a political problem, inasmuch as it is not merely political on a thematic level of how AI is used within our political institutions but in its very logic that is defined by putting forward values and seeking their realization.

To understand the practitioner's view regarding values, ethical guidelines and their realization requires taking the political nature of the context of the complex field of AI more seriously into account. Our claim is that political theory, precisely because it is focused on the links between norms and action, prescription and realization and justification and enforceability can help with understanding a further level of the emerging AI problem (Damnjanovic 2015, 76; Schippers 2020, 35). In other words, to understand on a practical level how to regulate AI one needs to consider the political context of regulation.

Therefore, the article aims to sketch a twofold problem that emerges at the intersection of AI and political theory. First, we demonstrate that the implications of approaching AI from a political theoretical perspective require that AI should be considered with a particular focus on the special nature of the political sphere. Here, we challenge the 'applied ethics' approach of those authors who regard AI as a mere regulatory problem. We claim that in discussing the regulation of weak AI, the context of political action and the specific conflictual nature of politics must always be taken into consideration either on the level of the emergence of AI within political practice or in the more abstract understanding of the general political context of AI. We address this as a problem of political realism: building on an analogy, we claim that just as the current mainstream in AI research conceives of guidelines as 'applied ethics', problems of relevance and efficacy should be reconsidered, just as realist political theory did to the relationship between normativity and political action. Second, we claim that the implications of AI for political theory could be beneficial for addressing some of the oldest political theoretical questions; for example, issues of peaceful coexistence, sovereignty or authority. Here, we claim that AGI and superintelligence are relevant tools for political theoretical inquiry.

The authors of this article are political theorists with no formal training in AI. But we are enthusiastic about both the nature of the political world and the possibility of political theorizing on the practical guidance of human conduct. We believe that our political theoretical contribution may help AI experts to understand the inherent limitations of regulating AI. At the same time, we see the current interest in AI as a great opportunity to understand the ordinary operation of the political world.

2. The realist approach

We take political realism as a framework for discussing the political context of AI guidelines. In this section of the paper, we briefly define political realism and highlight what we believe is an intrinsic link between the main concern of realist political theory and contemporary efforts to prescribe the development and functioning of AI.

Realism has always been a main general understanding of politics, even if during the latter part of the 20th century it became predominantly known as a paradigm in IR where it was described as politics focusing on power, interest or the use of force (Morgenthau 2005). Whereas these terms are not alien to contemporary realist theory, the focus of the authors belonging to this loose and heterogeneous 'movement' is broader. Recent realist resurgence was triggered by political philosophies which conceived of politics as the application of pre-existing ethical principles elaborated via abstract reasoning. Contemporary political realism holds that any normativity in politics, provided we want it to be effectual, must conform to the specificities of politics. That is, realist political theory is a countermovement against political philosophies that disregard the autonomy of politics and see it as nothing more than applied ethics.

Political realists do not deny that normativity is an ineliminable part of politics. In fact, contemporary realist theorists maintain that politics deserves its name as a form of legitimate rule due to its contrast to mere domination. Unlike mainstream political philosophy, however, realists do not think that abstract values such as justice could or should be considered before answering the 'first political question' of securing order as the condition of cooperation (Williams 2005, 3; Sleat 2018, 5–6). Order, security or legitimacy are values intrinsic to politics, therefore their realization conforms to the specificities of politics thereby providing actual normative standards for political action. By contrast, proponents of justice, equality and fairness fail to understand politics properly when they begin by elaborating abstract ideals and expect them to be applied to the sphere of politics. The same applies to ideals of transparency or explainability, along with some other values often discussed in the context of AI ethics. Moreover, the applied ethics approach can be regarded as an attempt 'to evade, displace or escape from politics' (Galston 2010, 386) against which realists, to become politically more relevant, demand the autonomy of politics and political thinking (Galston 2010, 408; Geuss 2008, 23; Sleat 2011, 471; Williams 2005, 3). Applied morality, as Bernard Williams calls it, is a mistaken way of conceiving normativity in politics (Williams 2005, 2).

Mainstream political philosophers, advocates of the applied ethics view, maintain that politics is just one of the spheres that has its own set of rules, yet the source of normativity for them remains abstract morality (Leader-Maynard and Worsnip 2018, 761). Discussions of several areas which have quite specific codes of conduct, ethical guidelines and regulations (such as medicine) are predominated with this approach. We argue that this approach fails to answer a crucial question: how to expect abstract values to regulate, justify and measure (political) behaviour without having the necessary means (enforceability) to become effectual in politics? Self-imposed guidelines fulfil their task only up to the point where they do not undermine the goals of their authors, or where the values incorporated into them do not conflict with each other. AI guidelines just because they are normatively attractive will not be binding without being enforceable. As realists tend to say, it is politics that provides collectively binding – and legitimate – decisions (Burelli 2020).

This leads us to a further insight about guidance conceived as a link between theory and practice. Realists criticize mainstream political philosophers due to the political irrelevance of their abstract morality. Realists claim that their approach is epistemologically better since they acknowledge that politics has its own set of rules and contextually embedded norms. From political relevance, realists proceed in two different directions. Some theorists maintain that political theory should elaborate measures, justification and critical standards for the political sphere. Realist political theory can give guidance to politics precisely because by acknowledging the autonomy of the political realm, political theory can become more relevant (Rossi and Sleat 2014, 689). Others, however, suggest that more relevant – realist – descriptions of politics mean that political theory should give up any aspirations to become action-guiding; it should remain a descriptive effort, and focus on what 'thinking politically' means (Horton 2017; Freeden 2018). It might be a reasonable compromise to acknowledge that theory can become action-guiding, but what practitioners actually 'follow' is 'contingent and, crucially, incidental' to the purposes of political theory (Horton 2017, 498). That is, it is unreasonable and unrealistic to expect that only because a guideline meets the abstract standards it becomes binding and relevant for practice.

As we have seen, realists do not claim that politics should be devoid of normativity. At the same time, realism is considered as a methodological innovation in political theory which, instead of putting forward specific proposals for political action, argues for a closer look on politics before defining the values to be realized (Rossi 2016; Jubb 2017). Although in what follows, we will make use of some of the substantive claims offered by realism, on the whole, our aim is to draw attention to politics both as the actual context of guiding and regulating AI and as the nature of these efforts that aim at bridging abstract norms and practice. That is, we do not suggest that the AI guidelines problem can only be dealt with by adopting realism. Realism, however, is a good way of showing the limits of guidance in a context where the need for binding decisions appears on the horizon.

3. The politics of AI: Enforceability and collectively binding decisions

In the context of current debates on regulating AI, politics becomes meaningful on two levels. First, the *AI in politics* problem concerns the use of artificial intelligence in politics. According to an instrumental understanding, AI needs to be regulated due to its profound (often worrying) impact on political practice from campaigning to administrative decision-making. Second, the *politics of AI* needs to be discussed to delineate the political nature of the problem of regulation, by which we mean that any attempt to achieve a normatively binding prescription for developing and applying AI is necessarily a political action. As for the former level, we take concerns about the proliferation of AI and the hopes regarding its regulation to be based on an idealistic understanding of politics. Taking the much-debated issue of interfering with voter behaviour as a brief example, regulation often comes down to attempts to exclude manipulation from political practice. As the Cambridge Analytica scandal or the Brexit referendum (Schippers 2020) exposed, political professionals rely heavily on machine learning hoping that the vast amount of voter information they gathered can be turned into meaningful data for segmentation, targeting and messaging at unprecedented levels of precision. Although it is worth noting that, especially in US electoral politics, computational tools are far from new and have been present since the 1950s (McKelvey 2021; Issenberg 2012), recent advances in AI precipitated worries about how technology undermines democracy. AI is seen to be weakening democracy by increasingly centralizing and controlling information and communication, creating fake identities, support and messages, as well as by altering the perceived political reality by reinforcing filter bubbles (Savaget 2019). In sum, the problem with AI and the reason why it should be regulated is that it provides politicians with more efficient means of manipulation whereas voters become more disempowered, and thereby accountability as a fundamental feature of democratic governance becomes void. However, more realistic authors have always been more cautious with the idea that voters are autonomous agents and that electoral politics without manipulation does exist (Schumpeter 1987; Körösényi 2010; Achen and Bartels 2016). Separating acceptable and unacceptable forms of influencing voter behaviour is a debate that is already political and cannot be solved by abstract ethical principles or self-regulation however carefully elaborated. Limiting political manipulation is not something beyond or before politics but a part of it, therefore politics seriously constrains the possibilities of how and to what effect the use of AI can be constrained. It is here where the AI in politics problem turns into the more abstract question of the politics of AI to which we turn now.

3.1. The role of the state

The problem of the state serves as a link between concerns about the use of AI in politics and the more abstract problem of the politics of AI. On the one hand, the state is evidently implied in the problems described above, yet it is hard to find in the discussion about AI ethics and guidelines. On the other hand, the general political nature of AI and AI regulation might be approached through the concept of the state, even if it is far from being the sole route a political theorist might follow. The absence of the concept of the state from the AI ethics discourse is somewhat understandable inasmuch as the topic of regulation through guidelines is focused on self-governance. As Hagendorff (2020, 100) remarks, ethical guidelines developed by companies and research institutions serve to discourage the creation of a 'truly binding legal framework'. Even if such strategic implications might not be general in all systems of AI ethics, enforceability is an open question that cannot be answered by completely disregarding states as key players in the field. The role of the state emerges on two levels.

First, as Hagendorff rightly observes, ethics – AI ethics included – cannot reinforce itself as it lacks the necessary – coercive, we might add – means (Hagendorff 2020, 99). The state is obviously the primary existing candidate for this role. However, considering the discussion on AI guidelines it seems that AI ethics – as a kind of applied ethics – expects that its principles will be followed merely due to their rightness, epistemological soundness, and normative attractivity. This, however, does not answer enforceability or a situation in which competing values are present. Understandably, any self-regulating effort will reflect the particular position of its author even if there are attempts to put forward universal norms as well. Although not necessarily motivated by selfishness, the proliferation of ethical guidelines results in a proliferation of particular positions and the possibility of conflicting values as well. Recent studies (Hagendorff 2020; Héder 2020) have revealed several overlaps but also differences between these documents. From this, it follows that beyond lacking the means to enforce a guideline, a further problem results from there being many possibly conflicting guidelines. Just as with ethical beliefs in a society, in the absence of an arbitrator, not only does implementation remain unsolved but conflict resolution as well. Whereas the first problem implies inefficacy, the plurality of guidelines might result in a disorder of unaligned particularities.

The other problem concerns the supranational level of enforceability. Evidently, even if binding guidelines exist on the national level enforced by states, most of the problems that required regulation in the first place remain unsolved between and above states. However, it is unlikely that a global solution might be adopted in the form of a supranational regulatory agency (Erdélyi and Goldsmith 2018). Any global attempt to regulate the development and functioning of AI will be just as efficient as any previous effort to make binding decisions on, for instance, climate change or global tech companies. On a global level, states are as inevitable as they are insufficient when it comes to regulating AI at least until the establishment of a global government which, from a realist point of view, seems to be highly unlikely.

3.2. Value transfer and enforceability

As the condition of transferring values, the problem of enforceability emerges on a more abstract level, revealing how AI guidance is political by nature regardless of any actual value or political content. The applied ethics approach in the field of AI guidelines has become increasingly nuanced as more and more regulative levels and methods have been differentiated. Christian's (2020) alignment problem, for example, addresses the questions of the methods and contents of value transfer from humans to (narrow) AI. Considering the emergence of superintelligence, Bostrom's value-loading problem also differentiates various ways of transferring human values into AI (Bostrom 2014). The discussion however remains focused on the interaction between normativity and technology while politics, if it appears at all, seems merely to be a disturbing factor to be neutralized. Criticizing Bostrom's account, however, Totschnig (2019, 916) emphasizes that the predominantly technological approach misreads the nature of control over a future superintelligence inasmuch as it should be considered as driven by the political dimension of self-interest. To avoid a warlike situation between a superintelligence and humans mutually considering each other as an existential threat, and achieving a peaceful coexistence, AI must not be antagonized by treating it as a tool or servant (Totschnig 2019). While Totschnig's realist implications are promising, in the end, the described relation between humans and AI becomes idealistically depoliticized and the seemingly political model fails to address the relationship between value transfer and enforceability.

Totschnig notes that the mutual existential threat that characterizes the warlike situation between AI and human agents lasts only until the AI begins to develop into superintelligence (Totschnig 2019). From that point, AI will have control over all the necessary means to transform the mutual threat into a one-sided vulnerability of humans. Certainly, if the superintelligence decides to get rid of humans, their political situation dissolves. The other option, however, namely the peaceful coexistence achieved by recognizing the self-interest of AI, surprisingly fails to develop an adequate account of politics and the political context of regulation. Contrary to what Totschnig says here, peaceful coexistence through recognition does not end the Hobbesian warlike situation but actually extends it towards politics. Let us remind ourselves that under the realist framework politics is more than successful domination; it is a legitimate form of rule that is not merely acknowledged just because there is no other viable option but understood as acceptable. A threatening superintelligence and acquiescent humans cannot have a political relationship. Both the value-alignment and the value-loading problem should be raised at this point. Considering that instead of mutual vulnerability humans are now disproportionally weaker, no guideline can be transferred to AIs based merely on the attractivity of its principles. The situation becomes political when coexistence with superintelligence exceeds mere acquiescence and resignation. That is, to be called properly political, any cooperation or order needs to be justified, however, justification emerges from within the very context of coexistence and the shifting power-relations between human and AI agents.

This extension of Totschnig's reading of Bostrom's approach is meant to be a model of how AI guidance can be conceived. Analogous to what we said about how the realist position reveals the boundaries within which actions and relations can be called political, superintelligence in the extended example above serves as the extreme case of guiding AI. Given that the wider AI problem is about the externalization and delegation of more and more human decisions to artificial intelligence (Chiodo 2020), no paradigmatic difference exists between narrow AI and strong AI or superintelligence, when human decision-makers are expected to develop guidelines which, present-day at least, relate to the development, use and functioning of AI as well. Following from our 'politicized' account of the context of AI guidelines, it might be concluded that the political relation is multi-layered, and encompasses human versus human, human versus AI, and possibly AI versus AI relationships, given the condition of enforceability and binding prescription emerging on the horizon.

4. Implications of AI to Political Theory and the concept of regulation

In the previous section, we demonstrated how a political theoretical perspective on artificial intelligence reveals its particular political nature, and we showed why it is mistaken to regard AI as a regulatory rather than a political problem. In this section, we flip the perspective and unpack the implications of addressing AI as a political problem, that is, we show what political theory can learn from discussions about the regulation of artificial intelligence. We argue that two fundamental implications follow from this perspective. One implication is that AI and claims to its regulation embody a compilation among the traditional problems of political theory. Therefore, the seemingly new fears around AI and its regulation are, in fact, well-known problems for political thought. The other implication involves claiming that AI – even in its strong form, such as artificial general intelligence or superintelligence – is a valuable tool for understanding the political realm if applied as a particular political theoretical methodology, precisely, in a thought experiment.

4.1. The impossibility of regulation: A classical problem

The increasingly popular topic of artificial intelligence may seem marginal from a political theoretical perspective. However, contemporary debates about some dimensions of AI are, in effect, political in the sense that they revive some classical dilemmas of political thought. The current discussion about regulating AI and the need for ethical guidelines – perhaps because of the urgency of such claims – is surrounded by an unsettling atmosphere, whereas regulation is a longstanding issue for political theory. Roughly speaking, politics is exactly about the problem of control. That is, now we show how current debates on AI guidelines reiterate earlier concerns of political theory.

The idea of an artificial entity and its inherent dangers to humanity has engaged ancient imagination, for example, in Greek mythology, in the form of Pandora, an artificial person created by Hephaestus (Pereira 2021). However, artificial intelligence has appeared in modern political thought as a distinctively political idea. Hobbes is rightfully known as 'the grandfather of AI' (Haugeland 1985, 23) for two reasons. First, he invented the idea of reasoning as computation, and second, he elaborated the idea of the application of an artificial person for politics (Hobbes 1651, 1655). In his reasoning, to overcome the brute reality of the state of nature, in which conflict is permanent, the 'unity of the multitude' brings into being the 'state,' the Leviathan, which is a 'fictional character', by authorizing a representative, who represents the state by acting in the name of it. This representative, the 'sovereign', is another artificial person (embodied by a natural person or an assembly) who also lacked any existence before the act of covenanting (Skinner 2018, 358). Therefore, the Hobbesian theory of social contract aims to not only argue for the desirability of political order, and thus, for the need for government, but also to offer a tool for justifying the legitimacy of rule (ibid. 360–361). While the Hobbesian conception of war of all against all in the state of nature refers to the dangers inherent in the absence of political order (as it was seen above concerning our expansion of Totschnig's reading of Bostrom's superintelligence), Hobbes's concerns for legitimate authority imply a different type of danger inherent in political life. The conditions for the right to rule – one of which states that only the sovereign is authorized by governance and the other that governance must aim for the preservation of life and health of the members of the commonwealth (ibid.) – indicate that these artificially created entities always include the potential to exceed the constraints set for them. In the Hobbesian framework, this means that for the artificial person of the state, for which the metaphor of Leviathan seems particularly apt in this regard, there is a permanent threat of seizure by someone without proper authorization. At the same time, there is also an indispensable danger that the artificial person of the sovereign is a counter to the common good. The fear of the inherent potential of overreaching the scope of the authority is more explicit in Locke's social contract theory that permits the withdrawal of obedience to the sovereign in case of abuses of power (Locke 1689).

It is clear, therefore, that the aspiration to restrict artificial entities emerged at the very moment when the idea of artificial intelligence emerged. Nonetheless, and more importantly, artificial intelligence as a political concept is not only interconnected with attempts to specify its limits but can be regarded as a mechanism for balancing two persistent dangers of the political realm: the extremes of the disorder of the state of nature and the tyrannical, illegitimate use of force or even terror. Therefore, the original idea of artificial intelligence as a political conception highlights the inherent fragility of *the political*. In sum, the state and the sovereign, as archetypes of artificial political entities can at the same time offer desirable solutions and severe challenges for political life.

Recalling the realist viewpoint of the political sphere, it seems that the only attainable goal is a *modus vivendi*, which resonates with the idea that an inherent characteristic of the political world is balancing the possibilities of two extremes. History of politics supports this more pessimistic view: occasionally, eruptions of civil war and failed states still embody the brute reality of the Hobbesian state of nature, while the existence of authoritarian and totalitarian dictatures altogether with hybrid regimes are eternal reminders of the impossibility to limit power in a once-and-for-all manner.

In light of this reality of the political world, new claims for the regulation of artificial intelligence, more specifically on weak AI, are less promising. Debates on the regulation of AI concentrate on the need to connect principles such as fairness, accountability, safety, sustainability, and social inclusion, among others, to AI governance (for a more exhaustive list, see Hagendorff 2020). Nevertheless, the most discussed issue is transparency, which is among the primary claims for several AI ethics guidelines released by different institutions and companies in the past few years.

The current boom in ethical guidelines for AI involves several criticisms concerning the effectiveness of such guidelines based on their potential to implement transparency and other claims effectively. This line of criticism can be divided into three types of argument. The first type challenges the AI guidelines on their extensive list of ethical claims based on their ineffectiveness. This type, which can be called 'tick-box criticism,' can be coupled with a proposal of some different approach, for example, virtue ethics (see Hagendorff 2020). The second type, which can be called 'double standard criticism', is more sceptical about the possibility of guiding AI and whether full transparency can be achieved at all. This criticism builds on the argument that it would be a double standard to call for higher transparency in AI compared to human decision tools and human reasoning (see Zerilli et al. 2019). The third type of criticism is more focused, what we call 'specificity criticism', and argues that current Artificial Intelligence Guidelines (AIGUs) are not specific to AI, but they are simply attempting to gain social control over technology. This criticism also demonstrates that transparency and explainability are claims that specifically concern AI because in such cases there is a possibility of the autonomy of AI. In that case, though, the double standard problem arises (see Héder 2020).

These criticisms imply that there is a profoundly political characteristic of AI. On the one hand, there is a relative autonomy inherent in AI that can be understood in a broader sense. It is impossible to regulate in every detail, something that can develop by itself. On the other hand, concerning the expert systems of weak AI, the double standard criticism and specificity criticism correctly acknowledged that it would be an unfair expectation to regulate the decision-making of artificial intelligence in domains where human decision-making cannot be entirely regulated likewise. However, contrary to the double standard criticism, we do not base our argument on the similarity between the obscurity of artificial decision tools and human cognitive processes. Instead, we build our argument on the political characteristic of AI. Using AI as a tool is similar to political authorization: although accountability is the main virtue in politics, it would be unrealistic to expect legislative, executive, or judicial officials to act 'perfectly'. We can only hope that they behave to the best of their knowledge, and while we usually hold them to account for significant breaches of their power, mostly, we authorize them because authorization is the only legitimate way to create order without slipping into a Hobbesian state of nature or a tyrannical regime.

4.2. Using AI as assistance for practical thinking

From a political theoretical perspective, the differences between weak and strong AI and the differences between claims for their regulation are not striking. Hence AI resonates with an inherent problem of political theory, what we called balancing between extremes; our scepticism towards full transparency concerns both weak and strong AI. Moreover, the more 'fictional' ideas of artificial general intelligence

(AGI) and superintelligence are also highly relevant for political theorizing. Political theoretical methodology frequently employs some fiction in the form of thought experiments and intuition pumps. These methods help us test our reasoning, build and destroy arguments, or explore our intuitions. Therefore, they can be used for different aims, and for this reason, they can lead to entirely different conclusions. In fact, the idea of a state of nature is a typical thought experiment, a mental visualization used by political theorists to justify their arguments on particular issues. Nevertheless, there are other methods involving fiction in political theory, such as idealistic or even utopian ideas about a just society and a just world or assumptions about perfectly reasonable individuals in situations of complete information. Also, there are dystopian ideas in political theory about the absence of order or a surveillance state.

From the perspective of political theory, therefore, the possibility of the emergence of AGI or superintelligence is not as marginal as for more technical discussions. AI has the potential to reveal the complexity and unpredictability of the political world and the role of human conduct in shaping the political world. Using AI as a thought experiment as Bostrom and other authors who engage themselves with the idea of strong AI sometimes do, reveals that claims for transparency and *a priori* determined rules can never be entirely enforced. Armstrong (2007), for example, elaborates a relatively universal solution to guide future superintelligence, still addresses several cases in which regulation may fail, and admits that in some of these scenarios, we are 'screwed'. Bostrom also addresses a broad range of potential problems concerning the transparency and regulation of a future superintelligence.

The idea of the emergence of multipolar general artificial intelligence is also a helpful tool to understand politics. The AI race (even in its weak form) is similar to other technological races in human history, involving crucial political challenges, such as the race for the fission bomb, fusion bomb, satellite launch, or the ICBM (Bostrom 2014). As Bostrom demonstrates, governments have always been seeking to gain control over such projects, which may provide them with a decisive strategic advantage (Bostrom 2014, chapter 5.). Such a scenario powerfully highlights the ineliminable nature of conflict in the political realm. However, if the emergence of AGI or superintelligence involves the dissolution of conflict from politics, it would automatically mean the obsolescence of humanity as we know it.

Concluding the section, we aimed to argue for the relevance of applying the idea of AI – both in its prevalent, weak and in its less discussed strong forms, such as AGI or superintelligence. The point of our argument was to show that there is nothing new in AI that would be unknown to the political world. The origins of the idea of AI from Hobbes's political theory to contemporary realist political theory implied that *the political* could not be entirely subjected to human oversight. Politics is precisely about balancing between the absence of order and terror. Besides there being no final solution that secures politics once and for all, there is no possibility to regulate artificial intelligence and secure all of its desired virtues and norms in advance. However, this is not a pessimistic conclusion that refuses any attempts to implement such norms; rather it is a confident argument for demonstrating that we already have a toolset – available in the political realm – for keeping AI under control.

5. Conclusion

The article argued for a (fundamentally realist) political theoretical approach to the problem of AI regulation. Our aim was to take one step back from the current debates on AI guidelines and to investigate the context in which the claims of regulation appear, and – as a result – to question the 'applied ethics' type of attempts that aim towards *a priori* laid-down rules for AI. In the article, first, we sketched the main characteristics of a realist view, then we demonstrated how this perspective highlights that AI is, primarily, a political rather than a regulatory or a technical problem. In doing so, we identified two problems: one is about the problem of *AI in politics* and the other one is what we called the *politics of AI* problem. Regarding the former one, we showed our concerns for the way AI transforms democratic politics. Concerning the latter one, we discussed the role of the state in the enforcement of AI regulation; while claiming that there is also a deeper problem of choosing, aligning, and loading values we want for AI. Finally, we addressed what current attention to AI can teach political theory. In this regard, we first demonstrated that the question of regulation is a classical and irresolvable problem of political thought, to which any attempts seem to be doomed to failure. Second, we showed that taking AI as a thought experiment may help us understand how our political world operates.

The article touched upon some further issues that should be considered not only from a (realist) political theoretical view, but from a broader scope of discussions as well. One question is about the arbitrariness of the values we seek to regulate and to be implemented in AI. While current debates are focusing on the problem of formulating AI in a way desirable for us (let us say, for humanity *per se*), there is a preceding problem of which values to choose and what to do when there are competing or even conflicting values. Although it could be justified that values such as transparency and human oversight are primary values from the perspective of political theory's general commitment to democratic values and participation, it can be argued that some directly emancipatory values such as fairness and solidarity in AI are just as important.

Finally, we revisit the criticism we applied in the article. Basically, we argued for a political realist approach due to its methodological innovations and we mentioned the usefulness of some of its substantial claims. However, there could be other approaches that criticize the so-called 'applied ethics' approach to the regulation of AI from a different perspective, and in fact, there are some attempts for a virtue ethics view (see Hagendorff 2020). We did not intend to exclude the appropriateness of such perspectives in substantive matters. Rather, we attempted to address the problem of AI from a broader perspective that takes politics seriously.

References

Armstrong, Stuart. *Chaining God. A qualitative approach to AI, trust and moral systems.* Unpublished manuscript. 2007.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford United Kingdom; New York, NY: Oxford University Press, 2014.

Burelli, Carlo. "Political Normativity and the Functional Autonomy of Politics." *European Journal of Political Theory*, 2020. https://doi.org/10.1177/1474885120918500

Chiodo, Simona. "The Greatest Epistemological Externalisation: Reflecting on the Puzzling Direction We Are Heading to through Algorithmic Automatisation." *AI & SOCIETY* 35, no. 2 (2020): 431–40. https://doi.org/10.1007/s00146-019-00905-y

Christian, Brian. *The Alignment Problem: Machine Learning and Human Values.* W. W. Norton and Company, 2020.

Damnjanović, Ivana. "*Polity* Without Politics? Artificial Intelligence Versus Democracy: Lessons From Neal Asher's Polity Universe." *Bulletin of Science, Technology & Society* 35, no. 3–4 (2015): 76–83. https://doi.org/10.1177/0270467615623877

Erdélyi, Olivia J., and Judy Goldsmith. "Regulating Artificial Intelligence: Proposal for a Global Solution." In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 95–101. New Orleans LA USA: ACM (2018). https://doi.org/10.1145/3278721.3278731

Freeden, Michael: "Political Realism: A Reality Check." In *Politics Recovered: Realist Thought in Theory and Practice*, edited by Sleat Matt, 344–68. New York: Columbia University Press, 2018. http://www.jstor.org/stable/10.7312/slea17528.18

Galston, William A. "Realism in Political Theory." *European Journal of Political Theory* 9, no. 4 (2010): 385–411. https://doi.org/10.1177/1474885110374001

Geuss, Raymond. *Philosophy and Real Politics*. Princeton: Princeton University Press, 2008.

Hagendorff, Thilo. "The Ethics of AI Ethics: An Evaluation of Guidelines." *Minds and Machines* 30, no. 1 (2020): 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hatamleh, Omar, and Tilesch George. *Betweenbrains: Taking back our AI Future.* GTPublishDrive. 2020.

Haugeland, John. *Artificial Intelligence: The Very Idea*. First MIT Press paperback edition. 1989Bradford Books. Cambridge, Mass.: MIT Press, 1985.

Héder Mihály. "A Criticism of AI Ethics Guidelines." *Információs Társadalom* 20, no. 4 (2020) https://doi.org/10.22503/inftars.XX.2020.4.5

Hobbes, Thomas. *De Corpore.* 1655.

Hobbes, Thomas. *Leviathan.* 1651.

Horton, John. "What Might It Mean for Political Theory to Be More "Realistic"?" *Philosophia* 45, no. 2 (2017): 487–501. https://doi.org/10.1007/s11406-016-9799-3

Issenberg, Sasha. *The Victory Lab: The Secret Science of Winning Campaigns*. 1st ed. New York: Crown, 2012.

Jubb, Robert. "Realism." In *Methods In Analytical Political Theory*, edited by Adrian Blau, 112–130. Cambridge: Cambridge University Press https://doi.org/10.1017/9781316162576.008

Körösényi András. "Stuck in Escher's staircase: Leadership, Manipulation and Democracy." *Osterreichische Zeitshrift Fur Politikwissenschaft* 39, no. 3 (2010): 289–302.

Leader Maynard, Jonathan, and Alex Worsnip. "Is There a Distinctively Political Normativity?" *Ethics* 128, no. 4 (2018): 756–87. https://doi.org/10.1086/697449

Locke, John. *Two Treatises of Government.* 1689.

McKelvey, Fenwick. "The Other Cambridge Analytics: Early "Artificial Intelligence" in American Political Science." In *The Cultural Life of Machine Learning*, edited by Jonathan Roberge and Michael Castelle, 117–42. Cham: Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-56286-1_4

Morgenthau, Hans J. *Politics among Nations: The Struggle for Power and Peace*. 7th ed. Boston: McGraw-Hill Higher Education, 2005.

Pereira, Luís Moniz. "The Carousel of Ethical Machinery." *AI & SOCIETY* 36, no. 1 (2021): 185–96. https://doi.org/10.1007/s00146-020-00994-0

Rossi, Enzo. "Can Realism Move Beyond a *Methodenstreit* ?" *Political Theory* 44, no. 3 (2016): 410–20. https://doi.org/10.1177/0090591715621507

Rossi, Enzo, and Matt Sleat. "Realism in Normative Political Theory: Realism in Normative Political Theory." *Philosophy Compass* 9, no. 10 (2014): 689–701. https://doi.org/10.1111/phc3.12148

Savaget, Paulo, Tulio Chiarini, and Steve Evans. "Empowering Political Participation through Artificial Intelligence." *Science and Public Policy* 46, no. 3 (2019): 369–80. https://doi.org/10.1093/scipol/scy064

Schippers, Birgit. "Artificial Intelligence and Democratic Politics." *Political Insight* 11, no. 1 (2020): 32–35. https://doi.org/10.1177/2041905820911746

Skinner, Quentin. *From Humanism to Hobbes. Studies in Rhetoric and Politics.* Cambridge, Cambridge University Press. 2018.

Sleat, Matt. "Liberal Realism: A Liberal Response to the Realist Critique.". *The Review of Politics* 73, no. 3 (2011): 469–96. https://doi.org/10.1017/S0034670511003457

Sleat, Matt, ed. *Politics Recovered: Realist Thought in Theory and Practice*. New York: Columbia University Press, 2018.

Smuha, Nathalie A. "From a "Race to AI" to a "Race to AI Regulation": Regulatory Competition for Artificial Intelligence." *Law, Innovation and Technology* 13, no. 1 (2021): 57–84. https://doi.org/10.1080/17579961.2021.1898300

Totschnig, Wolfhart. "The Problem of Superintelligence: Political, Not Technological." *AI & SOCIETY* 34, no. 4 (2019): 907–20. https://doi.org/10.1007/s00146-017-0753-0

Williams, Bernard. *In the Beginning Was the Deed: Realism and Moralism in Political Argument*. Princeton, N.J.: Princeton Univ. Press, 2005.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" *Philosophy & Technology* 32, no. 4 (2019): 661–83. https://doi.org/10.1007/s13347-018-0330-6