AI and Democracy: Balancing Risks and Opportunities to Maintain Meaningful Human Control

Patrik Bartak, Bram Delisse, Jelle Donders, Indra Gesink, Jaouad Hidayat, Vijay Vivekanandan.

This report was written for the "Where will AI fit in the democratic system?" case of the AI governance hackathon.

Introduction

The fast development of artificial intelligence (AI) could lead to solutions to humanity's greatest problems, but could also bring about extensive risks to our society. From a longtermist perspective, considering the place for transformative AI in our society is worth it. Even though there is some uncertainty in the speed of future AI development, improvement in capabilities is itself inevitable, and once a critical threshold is reached, the risks brought about are existential. The domain that tackles the risks imposed by AI within strategic frameworks is called AI Governance. In the last weekend of March, we got the opportunity to join the itch.io AI Governance Hackathon. In these 2 days, we deliberate on the following question: *Where, if anywhere, will AI fit, or fits already, in the democratic system?*

We first describe the democratic process, some basic terminology for artificial intelligence (AI), and meaningful human control, as background knowledge. Second, we analyze our research question in three parts: 1) we consider the conditions necessary to maintain democracy and meaningful human control, 2) we consider the failure points and potential risks posed by AI systems, and 3) we analyze the opportunities where AI can be integrated with the democratic system, using links to the previous two parts. Lastly, we include deep dives into two of the mentioned opportunities, before wrapping up with a brief discussion and acknowledgements.

Background

We start by discussing some background information on the subject. First we introduce the democratic system, then we introduce some foundational AI concepts and terminology that will be subsequently used, and finally we discuss the concept of meaningful human control.

The democratic system

An ideal decision process can be modeled as consisting of three stages. First, decision-makers (DMs) need to be informed. Second, the DMs form judgments or opinions regarding that information. Typically, this is done within a discussion or debate. Where the information concerns facts, this concerns what it is that people value. Ideally, these stages clearly precede the third and final stage where a decision is made.

In particular, a political process consists of many and interconnected cycles of decision-making, and can be modeled as a system with a throughput. Popular opinion informs politics and political decision-making as an input, which in turn influences (consequences and) popular opinion further in the future as an output. In closing off this subsection, we flesh out this process in more detail, in particular for the democratic system.

In a typical, representative democracy, citizens vote for representatives. The frequency by which they vote, and generally how the voting is organized, is also something they can themselves vote on. Other varieties of democracy for example include direct democracy, where each citizen is instead its own representative.

The ordered menu of representatives from which citizens in a representative democracy can choose is pre-created by political parties. These parties also create the political program of the party, which on a high level of abstraction prescribes the opinions that the representatives should have, and this certainty is used to inform the voter's choice. All this yields substantial power to political parties but on the other hand these usually also know an internal democratic process and anyone can start a new political party (even after being elected).

Elected representatives can take seats in parliament and collect political decision power. Often but not always a coalition is formed that occupies a majority of the seats, which is thus free to make decisions with majority rule. The representatives in this coalition propose individuals to take on responsibilities of governance. The government thus created is accountable to parliament, which checks them and ultimately sets the direction of policy.

Further, parliament and government are subdivided into many committees and subcommittees. A small third and neutral force mediates between parliament and government. They, for example, make the minutes and help to set the agenda for each meeting. They also check government reports in advance of parliament. Finally, the government employs many civil servants to actually develop the government's policies. These civil servants need to align with their employers, which can be subject to change with an election.

Execution of developed policy produces real-world effects, which can affect people, and/or provoke parliament, and thus inform voters and affect their decisions at the next election. The information voters receive is often mediated by the media, which can enhance information but also distort.

In closing, this whole ecosystem of collective decision-making relies on a large and diverse set of researchers, scientists and experts to provide factual information as a basis. Other organizations position themselves nearby. For example, think tanks, often driven by values, provide policy-makers with particular viewpoints.

Artificial intelligence

In order to narrow the scope we start by specifying the sort of AI that we consider in this paper. We also briefly introduce some terms used within AI safety to describe AI systems and the ways in which they may fail, namely robustness, alignment, interpretability, transparency, and algorithmic bias.

We start by defining agentic AI. This paradigm of AI from reinforcement learning is one of the most prevailing in the technical field of AI safety, and considers an agent embedded in some environment, perceiving this environment, taking actions to influence the state of the environment, and receiving some reward for that action. Addressing the problem of controlling a general form of optimizing agent is useful for long termist technical AI safety research, yet recent advancements in Large Language Models (LLM) have introduced an alternative form of AI that appears likely to be transformative in the near future, due to their impressive ability to compress knowledge and synthesize it in both natural language and other structured languages. These LLM's are likely to initially be deployed side by side with humans as "tool AI's", but may soon be integrated in most pieces of software. In this paper we therefore focus mainly on the role of such models in the near future. As mentioned before, the following terms are relevant:

Robustness broadly describes the ability of agents to achieve goals despite disturbances. Risks from insufficient robustness include: 1) distribution shift, where the distribution of data processed by the model changes over time, 2) vulnerability to sudden, unexpected, "long tail" events, and 3) vulnerability to interference by a malicious party through e.g. prompt injection, unauthorized access, or data poisoning.

Alignment is a crucial goal of AI safety research, and can be described in terms of outer and inner alignment. Outer misalignment describes the gap between our goals and the goals we specify to the system. Inner misalignment describes the gap between the goals we specify and the goals pursued by the system. The

main risk arising from this difference is the inability for the model to pursue our intended goals.

Interpretability consists of the ability to inspect and analyze the information contained in datasets and the goals of trained models. This allows us to evaluate the quality of AI systems and determine whether they can be deployed.

Transparency refers to the degree to which processes such as data collection, processing, model specification, training, testing, and monitoring are documented and accessible. Sufficient transparency is necessary in order to maintain a level of trust between AI and its users, as well as in order to provide auditability of the system.

Algorithmic bias refers to the tendency of AI systems to reflect human biases, whether present in data or the model training process. This can result in unjustly adverse outcomes for subpopulations in the data. Bias can originate from 1) the data generating process itself, 2) the sampling & processing of the data, or 3) the choice of model architecture & training procedure.

Meaningful human control

Meaningful human control expresses the need to maintain control and moral responsibility over otherwise autonomous systems. Originally this terminology was used in the military domain, but currently it is relevant much more broadly.

If, in general, we fail to retain control and responsibility over our AI systems, new and other forces than ourselves, whose goals we may not fully understand, will shape our world, thus undermining our own democratic influence and with that democracy itself.

How to maintain democracy?

In this section we first consider the necessary conditions for democracy, the absence of which can constitute threats to democracy. Second, we consider the maintenance of meaningful human control, within democracy and the context of developing AI technologies.

Necessary conditions for democracy

Various factors can be considered necessary or facilitating factors to maintain democracy. Their absence amounts to threats to the maintenance of democracy.

With all these necessary conditions satisfied, we might expect increased levels of shared understanding and cooperation, and engaged participation in democracy, without high levels of polarization.

- **Information** is primary in the democratic process. As presented in the background, participants need to be (properly) informed as the first step in a decision process. If and insofar informed untruthfully, the subsequent stages of judgment and decision formation will be misinformed and as a result potentially misdirected. If truthful, the information can be justifiably trusted.

  Sufficient truthfulness in the provided information is crucial in maintaining trust and thereby participation, which democracy depends on. Democratic participants need to remain sufficiently motivated if the democratic system is not to diminish.

  For example, deepfake videos - fake videos hard to distinguish from real ones, created by AI technology - and the like, widespread throughout society, could fundamentally turn our trust in information into doubt.

- **Prosperity**, or living standards that do not decline or stagnate too much, appears to be another necessary condition for democracy. If too unsatisfied, unmotivated democratic citizens tend to actively degrade their own democracy. Some degree of economic wealth thus seems necessary to maintain motivation required for the upkeep of democracy. Lack of wealth not in absolute terms but relative, i.e. in comparison, can have a similar effect, as well as cause absolute lack of wealth later on.

- **Justice** in terms of an absence of biases is another similar necessary condition. Its absence, in the form of blatant injustices, can spark such large-scale dissatisfaction that this could threaten the democratic system itself.

- **Equality** or distribution of power is, again, an adjacent necessary condition. If influence on the democratic system is too concentrated in the hands of a small number of actors, this threatens the democratic functioning. In the extreme, this is referred to as ``regulatory capture'', where the citizens no longer vote for the law of the(ir) land, but a small number of special interest groups (e.g. corporations) dictate instead.

- **Privacy** is necessary as leaked personal information could undermine trust. Personal information in the hands of others could be used against the person, exacerbating all of the other threats within this itemization.

- **Security** of attacks of military significance is another factor. If, in war, or under other direct threats to security, such as terrorism, democratic freedoms often are and/or need to be lifted. And, of course, a defeat in conflict can very well adversely affect a country's democracy.

- **Autonomy** of individual citizens is required for the expression of their personal values and beliefs within the democratic process. The absence of their autonomy undermines the integrity of this democratic process. An example of that would be a failure to resist psychological manipulation. This can be on the scale of one individual, or on a larger scale. For example a foreign attack, falling within the above category or often (and on purpose) just falling short of being of sufficient significance.

  On a larger scale the maintenance of a nation's democracy can also be viewed as synonymous with the maintenance of that nation's autonomy.

- **Deliberation,** and the means to do so, for example having time available, is another condition necessary for a well functioning democracy. In its absence, a decision-maker has no opportunity to form high-quality decisions.

  People are asked to fulfill multiple societal roles: work, familial care, citizenship. Each of these roles compete for time and other limited resources.

  For example AI displacing human labor could free up time for civic duties, and thus have a silver lining, if simultaneously the resulting lack of money would not require more time investments or lead to despair.

  The means to deliberate of course also relates to the previous condition of autonomy or the lack of psychological manipulation by another entity. Also this is relevant on the level of the individual as well as the nation.

- **Forward-looking** is a necessary condition not so much for democracy right now, but for its ability to detect and mitigate threats in the future, thus improving its ability to sustain itself into the future. Having, again, room to deliberate on problems that are further into the future instead of deliberate allows for ``slow democracy'' concerning larger existential risks to our potential future.

  A particular existential risk to our potential that we investigate here is unsafe or unaligned artificial intelligence. To mitigate it, it is necessary to maintain meaningful human control over artificial intelligence. We discuss this in more detail in the next section.

Maintaining meaningful control with AI

Considering in particular the implementation of AI into the democratic system, some of the above factors are of increased or decreased importance. Of particular importance is then maintaining meaningful human control, another

necessary condition for the maintenance of democracy with AI. With meaningful human control, humans can ensure future systems can be corrected or disabled in case they threaten our values.

This requires that humans retain final decision-power and steerability. Extreme race dynamics in the event of transformative AI might strongly incentivize differently. The short-term competitive benefits from undermining one's democracy by yielding meaningful human control to AI can be significant, yet we must instead consider the long-term impacts of these choices.

Risks posed by AI

In order to fairly critique the opportunities presented by AI,, it is useful to consider the points of failure for AI systems that are likely to be employed in such cases. We distinguish these points of failure into the following categories:

Data generating process

- Biased data generation - Data generated by the environment can include undesirable biases, such as against minority groups.
- Distribution drift - A changing environment distribution over time can result in a slow, difficult to notice degradation of the model performance, such as applying models to a changing economy.
- Extreme events - Rarity of extreme events can cause unforeseen (untested) scenarios, that in turn cause unforeseen model behavior..
- Positive feedback loops - Changing environments using models that depend on the environment can cause feedback loops that start to behave chaotically.

Data

- Data collection - Interference or bias in the collection of data can influence the goals and performance of the final model,
- Interpretability of data - Inability to inspect the contents of datasets at a detailed level prevents bias from being detected at an early stage.
- Data poisoning - Manipulation of data by third parties can influence the goals of the final model.
- Privacy violations - Data leaks can result in privacy violations and costs to society.

Model

- Mistrust in designers - Lack of trust in the designers and operators of the model can result in mistrust of the resulting model, undermining credibility of the system.
- Model interpretability - Inability to interpret model goals reduces the ability to deploy aligned models.
- Inability to align models - Inability to align model goals with intended goals can result in models pursuing goals that were not intended.
- Interference in operation of model - Interference by third parties can result in models being applied in unintended ways.

Emergent

- Dependence - Constant use of AI tools can result in a systematic dependence on them and a degradation of alternative methods.
- Public discontent - Disagreements and moral objections as to whether and where AI should be used can result in conflict.
- Centralization - Centralization of power through AI can result in weaker checks and balances, as well as information monopolies.
- Supranational tension - Conflict between countries/groups can be caused by the tension brought on by power imbalances through AI.

Opportunities in democracies with AI

In this section, we attempt to provide a list of possibilities where AI could play a disruptive role in the democratic process, in the context of bringing about a hopefully positive change in democracy. The concept of democracy can be broadly classified into 1) public sphere, 2) elections, and 3) administration. In each of these broadly classified domains, we present various opportunities, together with the applicable necessary conditions for a well-functioning democracy following it in parentheses. For each opportunity we explain the specific proposal, elaborate on the technology involved, and link it to both benefits to democracy, and potential risks.

Public Sphere

Enhanced Information Accessibility and Visualization (Information):

- Proposal: Utilize AI to analyze large amounts of data and provide summarized, user-friendly visualizations for easy public access to relevant information.
- Technical Insight: Implement natural language processing (NLP) and machine learning (ML) algorithms to extract key information and generate visualizations, such as graphs, charts, and maps, that enable citizens to understand complex information quickly.
- Benefits to Democracy: By making information more accessible and understandable, this proposal can promote transparency in government, encourage public participation in policymaking, and enable citizens to make informed decisions in elections.
- Potential Risks: Biased data generation and collection could result in discriminatory outcomes, and a lack of transparency in the AI system could lead to mistrust of the government. To mitigate these risks, the AI system should be designed with transparency and accountability in mind and undergo regular audits to detect and correct any biases. Additionally, stakeholders from diverse backgrounds should be involved in the design and implementation of the AI system to ensure inclusivity and representation.

Streamlined Legal Processes (Justice, Autonomy):

- Proposal: Implement AI tools such as IBM's Watson Legal and Lex Machina to automate repetitive and time-consuming legal tasks, such as document review and legal research, to provide efficient and cost-effective legal advice to clients.
- Technical Insight: Utilize NLP and predictive analytics models to extract and analyze legal data and documents, and provide legal advice based on the analysis.
- Benefits to Democracy: By streamlining legal processes and increasing access to legal resources for citizens, this proposal can improve the efficiency and effectiveness of legal systems and reduce costs for individuals and the government.

- Potential Risks: Over-reliance on AI tools can lead to a degradation of critical thinking and legal analysis, while biased data collection and analysis can result in discriminatory outcomes such as biased legal decisions or wrongful convictions.

Ethical Social Media Content (Privacy, Autonomy):

- Proposal: Implement AI tools, such as Jigsaw's Perspective API, to curate social media content based on ethical guidelines, promoting healthy and respectful online discourse.
- Technical Insight: Utilize NLP models to analyze social media content, identify harmful or unethical content, and provide personalized recommendations to users based on their interests and preferences, while respecting their privacy.
- Benefits to Democracy: By promoting ethical discourse and protecting privacy and autonomy for users, this proposal can help foster a healthy and informed public sphere, and encourage engagement in democratic processes.
- Potential Risks: The model's operation can be interfered with, resulting in unintended outcomes such as censorship or discrimination. Biased data collection and interpretation can perpetuate discriminatory practices and reinforce harmful stereotypes.

Inclusive Public Participation in Policymaking (Deliberation, Equality):

- Proposal: Implement AI tools such as Pol.is and Deliberatorium to facilitate public engagement in policymaking, allowing citizens to voice their opinions and engage with policymakers.
- Technical Insight: Utilize chatbots, opinion polling, and online forums to provide citizens with a platform to voice their opinions, while employing ML algorithms to analyze citizen feedback and provide personalized policy recommendations.
- Benefits to Democracy: By providing citizens with a platform to engage with policymakers, this proposal can promote democratic deliberation and equality, while improving policy outcomes by incorporating citizen feedback.

- Potential Risks: Interference from third parties, privacy violations, and centralization of such a system pose risks. Centralization in particular is an interesting risk to consider because such a system would require all citizens of a country to interact through such a channel in order to voice their opinion.

Objective Evaluation of Political Parties (Equality, Information):

- Proposal: Implement AI tools, such as AdVerifai and Factmata, to provide unbiased evaluations of political parties and incumbents, focusing on policy positions rather than personality traits.
- Technical Insight: Utilize NLP and ML models to analyze news articles and other media sources, and provide objective evaluations based on policy positions and voting records, while ensuring that the data sources used are diverse and representative of different perspectives.
- Benefits to Democracy: By providing citizens with unbiased evaluations and focusing on policy positions rather than personality traits, this proposal can promote equality and information, improve the quality of political discourse, and encourage informed and engaged citizenship.
- Potential Risks: Data poisoning, interference, and lack of transparency in the case of such a system could misrepresent the overall value preferences of the democratic system and result in undermining of the system.

Data-Driven Journalism (Information):

- Proposal: Implement AI tools, such as OpenAI's GPT models, to generate news articles and reports based on factual data and objective analysis.
- Technical Insight: Utilize large language models to analyze and interpret data, and generate unbiased news content. Personalized news content can also be provided to users based on their preferences and interests.
- Benefits to Democracy: By promoting transparency and reducing the impact of misinformation and fake news, this proposal can improve information dissemination and accessibility for citizens, and foster a more informed and engaged public sphere.
- Potential Risks: The misuse of AI-generated content by malicious actors can spread misinformation and propaganda. To mitigate these risks, the

AI system should be designed with transparency and accountability in mind, and regular audits should be conducted to detect and correct any biases. Furthermore, there should be a mechanism in place to verify the authenticity and reliability of the AI-generated content to prevent its misuse by malicious actors.

Elections

Encouraging Voter Participation (Deliberation, Autonomy):

- Proposal: Implement AI tools, such as VoterScience, to analyze voter data and promote personalized voter engagement, encouraging participation in democratic processes.
- Technical Insight: Utilize ML models to extract voter preferences and values, and provide personalized recommendations to voters, such as matching voters with candidates or issues that align with their interests and values.
- Benefits to Democracy: By increasing voter participation and engagement, this proposal can improve the quality of political discourse and promote democratic deliberation, resulting in better policy outcomes that align with the interests and values of citizens.
- Potential Risks: Biased data collection and interpretation can result in discriminatory outcomes, such as perpetuating systemic inequalities or reinforcing harmful stereotypes. Additionally, the over-reliance on AI can lead to a degradation of critical thinking and decision-making skills among voters. Misuses of the AI system could include attempts to manipulate voters by providing biased or false information or using the system to unfairly target certain groups of voters.

Trustworthy Fact Checking and Media Analysis (Information, Autonomy):

- Proposal: Implement AI tools, such as Full Fact's automated fact-checking system, to provide reliable and objective evaluations of news articles and other media sources, promoting transparency and accuracy in media.
- Technical Insight: Utilize NLP models to analyze the factual accuracy of media sources, and provide personalized recommendations to users based on their interests and preferences, while respecting their privacy.

- Benefits to Democracy: By providing reliable and accurate information, this proposal can reduce the impact of misinformation and fake news on public opinion, and promote a more informed and engaged citizenry.
- Potential Risks: The over-reliance on AI tools for fact-checking can lead to a degradation of critical thinking and independent analysis. Misuses of this AI tool could include its use to selectively censor or suppress information that does not align with certain political or ideological agendas.

Efficient Policy Information Summarization (Information, Autonomy):

- Proposal: Implement AI tools, such as Ayasdi's Policy Advisor, to summarize complex policy information for policymakers, providing concise and easily accessible summaries of policy documents.
- Technical Insight: Utilize NLP and ML models to analyze policy documents, extract key information, and generate summaries that are accurate, comprehensive, and unbiased.
- Benefits to Democracy: By streamlining the policy-making process and providing policymakers with concise and accessible information, this proposal can improve the efficiency and effectiveness of policymaking, reduce costs, and increase the accessibility of policy information for citizens.
- Potential Risks: An AI system trained on policy documents that contain biases against certain groups may generate summaries that reinforce those biases, leading to discriminatory policy outcomes.

Personalized Voting Assistance (Autonomy, Security):

- Proposal: Implement AI tools, such as Voterfied and Votemate, to facilitate advanced voting arrangements by analyzing voter data and providing personalized voting recommendations, such as voting reminders or candidate matching.
- Technical Insight: Utilize ML models, such as Bayesian networks and decision trees, to process large amounts of voter data and provide personalized recommendations to individual voters, while maintaining the security and privacy of their information.

- Benefits to Democracy: By increasing voter participation and engagement, this proposal can promote democratic deliberation, improve the accessibility of voting options for citizens, and encourage informed decision-making among voters.
- Potential Risks: A voting recommendation model that disproportionately recommends candidates from a particular demographic group could perpetuate systemic inequalities. Additionally, the over-reliance on AI can lead to a degradation of critical thinking and decision-making skills among voters, potentially resulting in an uninformed electorate.

Election Integrity Assurance (Transparency, Security):

- Proposal: Implement AI tools, such as Amazon's Rekognition, to prevent malpractice and manipulation in elections by detecting and preventing fraudulent activity such as voter impersonation and ballot stuffing.
- Technical Insight: Utilize ML models to analyze voting data and detect potential fraud or manipulation in real time. These models can use facial recognition technology to verify voter identities, detect anomalies in voting patterns, and flag potential threats to the integrity of the electoral process.
- Benefits to Democracy: By promoting transparency and fairness in elections and preventing fraudulent outcomes, this proposal can increase public trust in the electoral process and promote a healthy democratic society.
- Potential Risks: The use of AI in elections raises concerns about privacy and security. Moreover, there is a risk that political entities may misuse AI to manipulate voters or influence the election outcome. Measures must be taken to protect voter privacy and ensure that the system is secure against potential attacks.

Administration

Corruption and Fraud Detection (Transparency, Justice):

- Proposal: Implement AI tools, such as Palantir's Gotham, to detect patterns of corruption and fraudulent activity in government data, promoting transparency and accountability in government.

- Technical Insight: Utilize ML algorithms such as anomaly detection and network analysis to analyze government data and identify patterns of corruption or fraudulent activity, such as nepotism or embezzlement.
- Benefits to Democracy: By detecting and preventing fraudulent or illegal activity, this proposal can promote transparency and accountability in government, increase public trust in government institutions, and foster a culture of integrity in democratic systems.
- Potential Risks: Concerns about privacy and security may arise from the use of AI in government, as sensitive government data may be vulnerable to cyber-attacks or unauthorized access. Misuses of AI in government could also include the use of surveillance tools to monitor citizens, potentially violating privacy and civil liberties.

National System Optimization and Improvement (Prosperity):

- Proposal: Implement AI models, such as those developed by Accenture and KPMG, to optimize and improve national systems and models by analyzing data and providing recommendations for sustainability and prosperity.
- Technical Insight: Utilize ML algorithms to analyze complex data sets, such as energy consumption and urban planning, and generate predictions and recommendations for improvements and optimization.
- Benefits to Democracy: By improving the efficiency and effectiveness of national systems and models, this proposal can promote prosperity, sustainability, and the overall quality of life for citizens, while reducing costs for the government.
- Potential Risks: If the model is trained on biased data that reflects past discriminatory practices, it may fail to provide fair and equitable recommendations for optimization. Additionally, the use of AI in national systems and models raises concerns about privacy and security and the potential for malicious actors to exploit the system for their gain. To mitigate these risks, the AI system should be designed with transparency and accountability in mind, undergo regular audits to detect and correct any biases and be subject to robust privacy and security protocols.

Policy Impact Forecasting and Analysis (Information, Deliberation):

- Proposal: Implement AI tools, such as the Policy Simulation Library, to forecast and analyze the potential impacts of new policy proposals, providing data-driven recommendations to policymakers.
- Technical Insight: Utilize ML algorithms to analyze large amounts of data and simulate the potential outcomes of policy proposals, based on historical data and predictive analytics.
- Benefits to Democracy: By streamlining the policy-making process and increasing transparency and accountability, this proposal can improve the efficiency and effectiveness of democratic decision-making, reduce time and costs, and enable citizens to make informed decisions.
- Potential Risks: AI could be used to generate misleading or false data to support a particular policy proposal, or to automate decision-making without human oversight.

Strengthened National and Cybersecurity (Security, Autonomy):

- Proposal: Implement AI tools, such as Darktrace and Cylance, to enhance national security and cybersecurity by detecting and preventing cybersecurity threats, analyzing security data to identify potential risks and vulnerabilities, and improving emergency response and disaster management.
- Technical Insight: Utilize ML algorithms, such as deep learning and neural networks, to detect and respond to cyber threats in real time, and analyze large volumes of data to identify potential risks and vulnerabilities.
- Benefits to Democracy: By enhancing national security and cybersecurity, this proposal can help protect citizens from cyberattacks and ensure the integrity of democratic processes. Additionally, by promoting transparency and accountability in security and defense, this proposal can foster public trust and confidence in government institutions.
- Potential Risks: An AI system that is biased against certain groups could result in discriminatory practices in security and defense. Additionally, the use of AI in security and defense raises concerns about privacy and civil liberties, such as the potential for mass surveillance and the violation of citizens' privacy rights. Privacy and civil liberties should be

carefully considered in the design and implementation of the AI system, with appropriate safeguards put in place to protect citizens' rights.

Public Sentiment Analysis and Insights (Information, Deliberation):

- Proposal: Utilize AI tools such as IBM Watson and Google Cloud Natural Language API to analyze public opinion and sentiment from social media and other public data sources, providing data-driven insights into public opinion and sentiment.
- Technical Insight: Implement ML algorithms to extract and analyze text and provide insights into public opinion and sentiment, while ensuring the privacy of individuals is protected.
- Benefits to Democracy: By providing data-driven insights into public opinion and sentiment, this proposal can improve transparency and accountability in policymaking, promote democratic deliberation, and increase public engagement in democratic processes.
- Potential Risks: If a model only considers data from one demographic, it may not accurately represent the views of the wider population. Additionally, the use of AI in public opinion and sentiment analysis raises concerns about privacy and civil liberties, as individuals' data may be collected and analyzed without their consent. To mitigate these risks, the privacy of individuals must be respected, and data collection and analysis should only be conducted with appropriate consent and ethical considerations.

Advanced Emergency Response Management (Security, Autonomy):

- Proposal: Implement AI tools, such as IBM's Watson Decision Platform for Emergency Management, to improve emergency response and disaster management.
- Technical Insight: Utilize ML algorithms, such as deep neural networks, to analyze emergency response data, identify patterns and trends, and provide data-driven recommendations for disaster response and management.
- Benefits to Democracy: By improving emergency response times and effectiveness, this proposal can increase public safety and security and

promote transparency and accountability in emergency response and disaster management, ultimately saving lives and reducing damage.

- Potential Risks: If the AI system is trained on data that disproportionately represent certain demographic groups, it could lead to biased decision-making in emergency response situations. Additionally, the use of AI in emergency response raises concerns about the transparency and accountability of decision-making processes, potentially leading to mistrust in the government or emergency response organizations.

Data-Driven Law and Justice Reform (Justice, Autonomy, Forward-looking):

- Proposal: Implement AI tools, such as COMPAS, to analyze legal data and provide data-driven recommendations for law and justice reform, improving the efficiency and effectiveness of legal systems.
- Technical Insight: Utilize ML algorithms to extract and analyze legal data, and provide recommendations for improving judicial decision-making, reducing sentencing disparities, and increasing fairness and accountability in law and justice.
- Benefits to Democracy: By promoting transparency and accountability in legal decision-making and reducing the impact of bias and discrimination on legal outcomes, this proposal can increase trust in legal systems and promote democracy.
- Potential Risks: The use of facial recognition technology in law enforcement has been criticized for perpetuating racial bias. Additionally, the use of AI in law and justice raises concerns about privacy and civil liberties, such as the potential for mass surveillance or violation of due process rights.

Conflict Resolution and Diplomacy (Deliberation, Forward-looking):

- Proposal: Implement AI tools, such as IBM's Watson Discovery and Palantir, to improve conflict resolution and diplomacy by analyzing diplomatic data and providing data-driven recommendations for peacekeeping efforts.
- Technical Insight: Utilize ML models, such as neural networks and decision trees, to analyze large amounts of diplomatic data, detect

patterns and relationships, and provide informed recommendations for conflict resolution.

- Benefits to Democracy: By increasing the effectiveness and transparency of conflict resolution and peacekeeping efforts, this proposal can promote accountability in diplomatic decision-making, increase security, and foster a more peaceful and stable global community.
- Potential Risks: Malicious actors could potentially misuse the AI system to promote their agenda or perpetuate conflict.

Deep dives

In this section, we consider two interesting opportunities in depth to integrate artificial intelligence in democracy. The study of the other opportunities is a task left for the reader but we hope that these two comprehensive examples make for a good start for it.

Deep dive - Encouraging Voter Participation

With regards to "Encouraging Voter Participation", we see AI as a potential facilitator towards engaging citizens in empowering democracy through their right to vote in elections. This applies specifically to the section of society that is not too involved in political and democratic processes, lacking the political literacy in terms of candidature parties, their ideologies, and the performance of various candidates/parties in their respective domains.

This takes the shape of a facilitator/tool accessible by all members of the public sphere (ex. as an app, dashboard or consultant), based on requirements of the end user. This can be elaborated as follows:

For citizens with no prior political/candidate knowledge, but the will to use their vote to make a difference in the values they believe in

It is a surprising yet factual argument that the voter turnouts have been undergoing a general global decline in democracies, as demonstrated in Fig. 1 and Fig. 2. While the reasons can be very nation-specific and citizen sentiment-based, there is general consensus that large sections of the society do not actively engage in keeping up with political affairs on a continual basis. It has also been studied that voter turnout and engagements are directly proportional to important elections, such as those that relate to the long-term political future of the country, and when the country faces fewer external constraints on policy making [3]. The effects of this ongoing trend can be catastrophic for democracy, as the votes of the public sphere is central to

keeping the balance of power, with the democratic foundation being built on the principle of "Government of the people, by the people, for the people".

AI as a facilitator could bring about a change through personalized and interactive engagements with the public sphere, increasing the knowledge and motivation to vote and make a difference. It could act as a tool that takes in a prospective voter's inputs with regards to the values he/she most believes in, the causes that connect the most with their morals, and the ethical values they abide by in their day to day lives. Based on this input, the AI can analyze the voter's ideologies for a better society and give feedback on the candidature profiles that most relate to the causes the voter believes in. While this can be looked at from an AI-based recommendation point of view, this does not

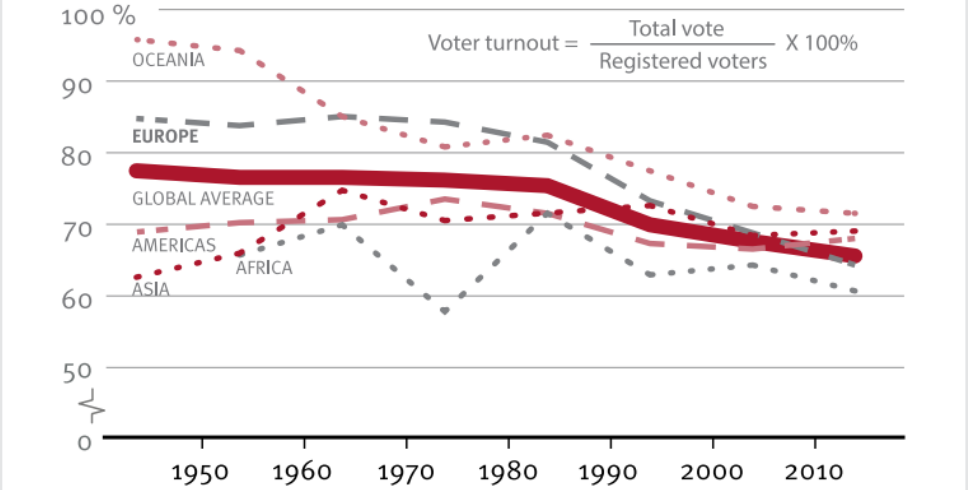
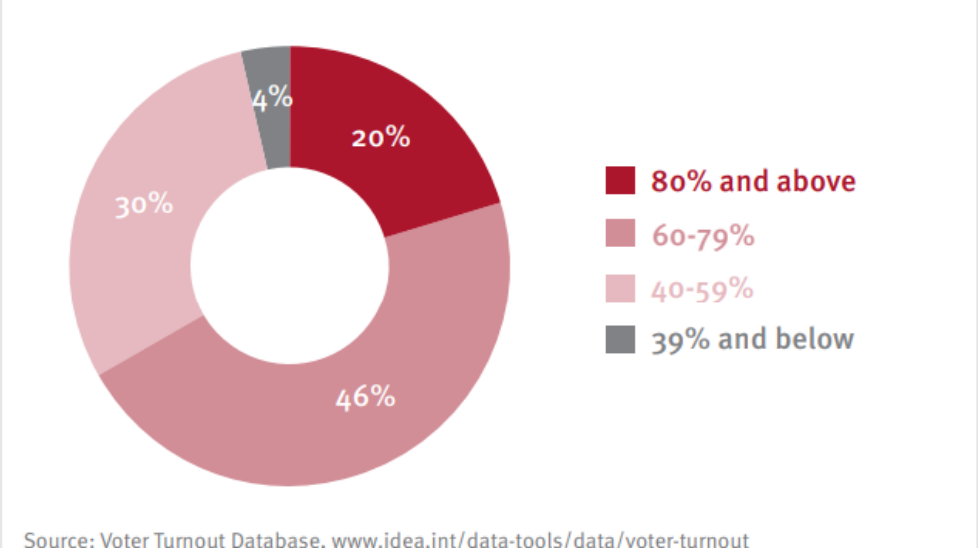


Figure 1: Global voter turnout over time.



completely hold true as 1) the tool does tend to show all candidature profiles though there is a sorting based-on personal connection, while the final decision-making still lies in the hand of the voter, 2) It is a better alternative to voting randomly based on no knowledge, or not voting at all, and 3) it brings about a personal touch to improving the society - a combination of individual and societal welfare.

Figure 2: Global voter turnout by country.

For citizens with moderate to expert political knowledge, requiring summarized insights into candidature profiles

A lot of things happen during a government's tenure and elections are often crucial and critical points of decision-making. It is a complicated affair for members of the public sphere (as voters or even media) to account for every detail before jumping into conclusive comparisons and decision-making. With the power of AI's data crunching and analyzation capabilities, this could indeed be a revolutionary possibility to effectively compare and analyze candidature profiles based on performance and fulfilled demands, including complete transparency to personal records aligning towards any criminal offenses and scandals. It could also be possibly used to compare current and projected performances of candidates if elected into power, based on previous elections and tenure data. Combined with an information loop on debates within the public sphere based on the legitimacy and validity of points put forward, such AI-powered summarization and analyzation platforms could provide the complete picture to a voter looking to make a change with their right to vote.

While the potential integrations and the advantages that AI brings to the table in this regard has been discussed, the potential risks have to also be considered. The biggest risk in this regard would be tampering of training data, and the potential psychological manipulation of voters to vote for an actor with malicious intent. The need to mitigate this is primary, through complete transparency of how the AI model reaches a specific decision based on inputs, and also a quantitative/qualitative measure of truthfulness of these AI models [4].

Deep dive - Data-Driven Law and Justice Reform

AI can play a pivotal role in the law and justice system from a couple of viewpoints. One could be as an advisor/analyzer in the domain of legislation, whereas the other could be as an assistant in the justice provision process. The two viewpoints will be elaborated as follows:

AI as a legislative advisor/analyzer

The parliament and legal systems are accountable for the review and conduct of the society based on the principles, laws and values drafted as part of the constitution and legislative frameworks. As the society develops and goes through the scale of time, it is probable that the initially drafted laws and abidance frameworks do not fully hold true anymore [2], and that there is a requirement to modify them or totally remove them. This also includes unintended but possible biases embedded, that may embed partialities to specific groups. The process of modifications happens through policy submissions, extensive reviews and majority approvals in the parliament houses. While this may be one dimension, the other dimension involves the interpretation of established and the manipulation of these laws by actors of malicious intent, to get away with unethical practices that have been committed for personal benefit. The entire process of policy reviews, debating policy changes and identifying re-interpretation/loopholes/mis-interpretations in existing law and order is such a tedious process. Let alone tedious effort, but manipulation of laws by malicious actors have caused entities (like banks) and governments losses in the scale of millions, while other cases have had these actors commit serious intentional crimes towards the society with proportionally lesser scale punishments.

AI could be revolutionary in this domain if employed to review the constitution, the legal frameworks, the database of recorded cases, crimes and investigations, and the case studies that have been conducted with regards to special cases that occurred as a ripple effect of manipulation, mis-interpretation or non-updation of existing legal frameworks. Based on the analysis, the AI would be expected to throw insights into possible loopholes, project possible consequences, and propose better policies based on previous experiences. This would lead to a very efficient legislative framework, ensuring greater possibilities to maintain fair and unbiased law and order in the society. An illustration summarizing the incorporation of AI in this process can be seen below.

AI as an assistant to the judgment process

This could be an interesting use case for AI as legal assistants to the judges mediating a court case. The AI could provide real time sentiment analysis checks of the convicted parties, while also being a tool for fact/information/background checks. Moreover, the AI could also be a tool for analyzing the possibilities that might have happened and the subsequent consequences/nuances of the case, especially complicated ones such as criminal cases. Having been possibly trained on laws and legal frameworks in place, the AI could also ease the study process for lawyers in terms of the applicable legislations and laws in the case

for their clients, with some progress already being made in this domain as stated in [1].

While the potential integrations and the advantages that AI brings to the table in this regard has been discussed, the potential risks have to also be considered. The core point to address in this regard would be human control over decision making, steerability and transparency regarding the decision outputs of the model, as the plausible risks entail data manipulation, and the undermining of the credibility of the legal system due to enhanced unexpected performance (where the flaws and loophole found cannot be logically justified).

Discussion

In this report we considered where artificial intelligence will fit in the democratic system. After considering background knowledge of the democratic system, artificial intelligence and meaningful human control, we considered how to maintain democracy. Specifically, we considered necessary conditions for the maintenance of democracy, including meaningful human control over artificial intelligence. Second, we considered risks posed by artificial intelligence (AI). Third we analyzed several opportunities for use of AI in democracies, including their potential benefits and risks and connection to necessary conditions for democracy. Last, we analyzed two of them in more depth.

Of course there are limitations to how much one can consider in the space of a single weekend. Further reflection and investigation is likely fruitful in many directions. It is unlikely that all of the 19 opportunities identified have equal potential, and each would benefit from a more thorough investigation of their implementation in society, and the magnitude of each of the identified potential failures. Nevertheless we believe that our report can be the start of further reflection on where AI will fit in the democratic system. We hope not to have provided any final answers, but to – so as to enhance democratic participation – spark further reflection on these topics and their significance for our future.

Acknowledgements: We thank the Delft AI Safety Initiative (DAISI) for locally hosting this AI Alignment Jam, organized by Apart Research. Thanks to the DAISI organizers - Jaouad Hidayat, Patrik Bartak and Rauno Arike - and Apart

Research for the organization. We wish to thank organizer Rauno Arike as well for helping us with this project.

References

1. Bar exam score shows AI can keep up with 'human lawyers,' researchers say. Available at https://www.reuters.com/technology/bar-exam-score-shows-ai-can-keep-up-with-human-lawyers-researchers-say-2023-03-15/

2. The irrelevance of the Netherlands constitution, and the impossibility of changing it. Available at https://www.cairn.info/revue-interdisciplinaire-d-etudes-juridiques-2016-2-page-207.html

3. Voter turnout trends around the world - idea. Available at https://www.idea.int/sites/default/files/publications/voter-turnout-trends-around-the-world.pdf

4. Technical AI Ethics - Truthful QA. Available at https://arxiv.org/abs/2109.07958