AI & SOCIETY
https://doi.org/10.1007/s00146-024-02104-w

OPEN FORUM

Knowledge and support for AI in the public sector: a deliberative poll experiment

Sveinung Arnesen[1] · Troy Saghaug Broderstad[2] · James S. Fishkin[3] · Mikael Poul Johannesson[1] · Alice Siu[3]

Received: 19 February 2024 / Accepted: 9 October 2024

Abstract
We are on the verge of a revolution in public sector decision-making processes, where computers will take over many of the governance tasks previously assigned to human bureaucrats. Governance decisions based on algorithmic information processing are increasing in numbers and scope, contributing to decisions that impact the lives of individual citizens. While significant attention in the recent few years has been devoted to normative discussions on fairness, accountability, and transparency related to algorithmic decision-making based on artificial intelligence, less is known about citizens' considered views on this issue. To put society in-the-loop, a Deliberative Poll was thus carried out on the topic of using artificial intelligence in the public sector, as a form of in-depth public consultation. The three use cases that were selected for deliberation were refugee reallocation, a welfare-to-work program, and parole. A key finding was that after having acquired more knowledge about the concrete use cases, participants were overall more supportive of using artificial intelligence in the decision processes. The event was set up with a pretest/post-test control group experimental design, and as such, the results offer experimental evidence to extant observational studies showing positive associations between knowledge and support for using artificial intelligence.

Keywords Deliberation · AI · Experiment

1 Introduction

The use of artificial intelligence (AI) and machine learning (ML) is on the rise in the public sector.[1] A report published by the European Commission in 2022 identified 686 public sector AI use cases in its member states plus some other European countries, most of which were based on machine learning (Noordt et al. 2022).[2] Several AI solutions are already deployed and in use. Determining parole, reallocating refugees, and deciding welfare eligibility are just three examples where AI systems are being used or under development. The majority is still in the pilot or development phase, so the scope and scale of AI use are likely to widen rapidly.

These computational advances combined with the growing availability of data raise novel governance challenges with respect to discrimination, fairness, and transparency in AI decision-making (Kroll 2015). With it, there is an

✉ Sveinung Arnesen
sarn@norceresearch.no

Troy Saghaug Broderstad
troy.s.broderstad@uit.no

James S. Fishkin
jfishkin@stanford.edu

Mikael Poul Johannesson
mikj@norceresearch.no

Alice Siu
asiu@stanford.edu

[1] NORCE Norwegian Research Centre, Bergen, Norway

[2] UiT, The Arctic University of Norway/NORCE Norwegian Research Centre, Tromsø, Norway

[3] Stanford University, Stanford, USA

[1] To ensure readability, we use'AI','AI/ML','Algorithmic' interchangeably. The difference in wording refer to the same concept of decisions made by algorithms. In the literature, these terms are used interchangeably and may sometimes refer to different concepts. For instance, automated decision-making systems (ADMs) may sometimes refer to a basic decision tree where humans program the criteria for how the automated decision is researched. In this article, we disregard this type of automated decision-making and focus solely on decisions made by a computer or algorithm that can synthesize and infer information based on data, without human intervention.

[2] The dataset of the use cases is available at the European Commission's Joint Research Data Centre Catalogue: https://data.jrc.ec.europa.eu/dataset/7342ea15-fd4f-4184-9603-98bd87d8239a.

Published online: 27 October 2024

increasing need—and demand—for empirical research on specific AI use cases in the public sector (Zuiderwijk, Chen, and Salem 2021). In democracies, input from citizens is critical to developing legitimate AI systems, not least in the public sector which has obligations to make sure its use of AI adheres to democratic principles. A pressing concern is thus how to—in the words of Rahwan (2018)—put society in-the-loop when developing and implementing AI tools in public administration. Our solution is to conduct a Deliberative Poll (Fishkin 2019), where we invite a representative sample of the population and provide them with the time and resources to learn about the topic and discuss trade-offs and ethical considerations about the use of AI in the public sector.

Our work offers several contributions. First, we provide a framework for how to increase AI knowledge among citizens and involve them in AI policy making—which includes not only acquisition of information regarding AI/ML, but also discussion with peers and interactions with subject-matter experts. Second, we implement the Deliberative Poll ($N=207$) with an experimental research design, which enables us to plausibly identify the causal effect of increasing knowledge for those who participated. Participants in the treatment group were given the opportunity to read balanced briefing materials on AI, discuss with peers in small-group sessions, and question experts on the area. Third, we consider multiple aspects of attitudes towards AI, including concrete examples being worked on. After a full day of deliberation, participants were surveyed on their self-reported knowledge about AI and about their attitudes towards using AI when deciding on reallocation of refugees, parole for inmates, and dialogue meeting with citizens on sick leave.

Extant literature suggests that citizens with more knowledge are more supportive of AI, however, the evidence to date has been correlational (e.g., Thurman et al. 2019; Logg et al. 2019; Zhang and Dafoe 2019; Araujo et al. 2020; Zhang 2021; Starke et al. 2022). Our experimental results show that increasing knowledge about AI/ML indeed makes citizens more supportive of AI in our three public sector use cases. Participants in the treatment group reported a significant increase in knowledge about AI/ML after the deliberation event, compared with the control group. In addition, their knowledge about the specific tasks where AI potentially can be used increased (refugee settlement and welfare-to-work). With the increase in knowledge, the participants in the treatment group also became more positive towards the use of AI when making decisions in these areas.

This paper proceeds as follows. First, we discuss why we should care about what citizens think about using AI in decision processes in the public sector. We review existing public opinion literature on the topic and identify the current challenge that citizens do not know or care much about the topic. We argue for the need to know the public opinion when they have had the chance to gain knowledge about the topic, and that one way of achieving this is to conduct a Deliberative Poll. Second, we present the design, fielding, and empirical results of a Deliberative Poll conducted in Norway on the topic of AI in the public sector. The paper concludes with a discussion of limitations and implications of the results.

2 Keeping society in-the-loop

2.1 Why citizen perceptions matter

While there are obvious efficiency benefits, challenges also arise with respect to perceptions of discrimination, fairness, accountability, and transparency in AI-based decision-making. Scholars highlight concerns for the potential of encoding discrimination in automated decisions, where discrimination may be the inadvertent outcome of the way big data technologies are structured and used (Barocas and Selbst 2016; Pasquale 2015). They fear a black box society where people's destinies are decided by opaque, imprecise, and discriminatory automatic decision makers. Policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of artificial intelligence and machine learning, calling for further technical research into the dangers of inadvertently encoding bias into automated decisions (Kroll 2015). The intentions of introducing AI into decision procedures may be benevolent, but several potential pitfalls loom, with consequences that ultimately may put the legitimacy of government institutions in jeopardy.

Government authorities depend upon being perceived as treating citizens fairly, and algorithms based on AI are not immune from the fundamental problem of discrimination, in which negative and baseless assumptions congeal into prejudice. The use of AI in the public sector is a special case because, unlike private companies, public authorities must conform to principles of democratically legitimate decision-making ((Beckman, Hultin Rosenberg, and Jebari 2022), see also Wirtz et al. (2020)). Non-elected officials with coercive powers such as police, prosecutors, and bureaucrats vary markedly in the extent to which citizens view their actions as legitimate (Dickson et al. 2015), and their legitimacy perceptions are influenced by to what extent citizens view them as trustworthy, transparent, reliable, impartial, uncorrupted, and competent government institutions (Tyler 2021; Rothstein and Teorell 2008). The introduction of AI governance can influence all these attributes, for better or worse.

The rise of AI in the public sector has thus brought recognition that it may affect the basic relations between citizens and their government. Hence, it is important that the development and use of AI in this domain is aligned with society's values. In theory, policy making could proceed without any public consultation, but then the public would be left in the dark and the values and concerns that the elites and experts bring might fail to address the public's concerns. Several investigations on citizens' perspectives have been undertaken in recent years to uncover how and where (see e.g., (De Fine Licht and De Fine Licht 2020; Binns 2018; Arnesen and Johannesson 2022; Grgic-Hlaca et al. 2018). It is fair to say that a general distaste for AI in the political realm is observed (König 2022), in particular for fully automated AI decisions (Starke and Lünich 2020; Waldman and Martin 2022). Also, the more important the topic, the more skeptical citizens become in giving such tools decision influence.

That being noted, an ongoing challenge with surveying the general population on their attitudes towards using AI specifically in the public sector is that the topic is low in salience and largely unknown to ordinary citizens. For example, a 2019 German study showed that almost half of the population did not know what an algorithm was (Grzymek and Puntschuh 2019). In 2023, four out of ten citizens of Norway reported little or no knowledge at all about AI/ML (down from six out of ten in 2021, see S.I. section B). A British survey displayed how people's awareness varied greatly across the different technologies they were asked about, where public awareness was lowest for less visible technologies, such as AI for assessing eligibility for welfare or risk in healthcare outcomes (Kantar 2023). Awareness of AI in general rose across the world with the release of large language models like ChatGPT, yet any discussion about how AI can and should be used in the public sector is still not receiving much attention. There is, therefore, a risk that opinion polls measure'placebo attitudes', where participants give an answer without actually having an opinion about the question (Luskin et al. 2002). Or, at least, citizens' attitudes on this topic are less well informed compared to attitudes they have on other topics. This is likely to change with time, but ideally, we would want to know what citizens think (or would think if they were well informed) before the AI tools are implemented, and not after.

Such early thematization of potential AI use cases is in line with the advocates of anticipatory governance (Fuerth 2011), who argue for the benefits of exploring and assembling current values, knowledge, and plausible scenarios to travel into the future with more rather than less reflexive governance capacity (Guston 2014). In accordance with their request for new types of institutions capable of addressing the dynamics of innovation and the societal impacts brought about by the latest scientific and technical advances, our aim with this study, therefore, is to experimentally induce a representative sample of the population with knowledge about AI, and then survey their (change in) attitudes. What will the people think after they have had the chance to become informed through reading about the topic, deliberating with fellow citizens, and questioning balanced panels of experts? How do these views compare with their pre-deliberation opinions? Our approach combines the aim of maintaining a representative sample of the target population while at the same time provide participants in the study the resources needed to acquire the necessary knowledge to form a considered opinion.

2.2 The impact of knowledge on citizen perceptions of AI/ML

There is a growing amount of evidence on the impact of knowledge and experience on public opinion towards AI/ML. Few studies, if any, focus exclusively on the impact of knowledge or experience, but many compare the attitudes of those with different levels of knowledge or experience (Zhang 2021; Starke et al. 2022). Most studies find that those with more knowledge or experience are more supportive of AI (e.g., Thurman et al. 2019; Logg et al. 2019; Zhang and Dafoe 2019; Araujo et al. 2020; Zhang 2021; Starke et al. 2022). Some results suggest that those with the most knowledge are again less supportive, suggesting a U-shaped relationship (Zhang 2021; cf. Lee and Baykal 2017; Zhang and Dafoe 2019). There are also more nuanced aspects of how knowledge or experience may impact opinion. For instance, Lee and Baykal (2017) and Saha et al. (2020) find that those with more technical understanding of fairness metrics tend to evaluate algorithmic decisions as less fair; Wang, Harper, and Zhu (2020) find that those with higher education tend to have more stable fairness perceptions regarding algorithmic decisions regardless of whether they personally benefit from them.

The extant literature leads us to the following hypothesis:

Hypothesis: More knowledge about AI/ML will increase the support for using AI/ML

However, there are many limitations to the current evidence, suggesting the hypothesis just as well could have been presented as a research question. First, how "knowledge or experience" is defined varies: Both math or computer programming skills (Logg et al. 2019; Lee and Baykal 2017), education (Thurman et al. 2019; Zhang and Dafoe 2019; Wang, Harper, and Zhu 2020), occupation (Zhang et al. 2021), and participants' self-assessed knowledge (Arnesen and Johannesson 2022; Araujo et al. 2020) have been used in previous observational studies. These can represent vastly different aspects of having knowledge, induce different impacts on opinion, and vary in terms of how relevant they actually are for understanding public opinion in this case. Second, no studies have tried to plausibly identify a causal effect of increasing knowledge. Comparing subgroups of participants provides interesting and important evidence, but various selection effects make it difficult to separate the impact of knowledge from other aspects that drives whom currently has such knowledge. To ascertain whether and how public opinion would change if citizens had more knowledge, we need to identify its unique effect on public opinion. As we discuss below, one way of increasing citizens' knowledge is through deliberation.

2.3 Knowledge through deliberation

The ideal of deliberative democracy is that citizens come together and express their views, listen to the opinions and considerations of their fellow citizens, reflect on them, and become wiser individually and as a group. An established definition of deliberation is'mutual communication on matters of common concern whereby participants weigh relevant considerations to inform conclusions regarding forms of action' (Bächtiger et al. 2018). The 'relevant considerations' should include competing arguments for and against the policy proposals under discussion.

As argued by democratic theorists, successful deliberation increases citizens' insight in complex cases and policy questions (Mansbridge et al. 2012). Increased knowledge is a main expected outcome of deliberation where citizens, under the right conditions, will increase their competence on the issue and reach the overall best collective decision. Also known as epistemic fruitfulness, it is a core criterion for successful deliberation (Estlund 2009; Landemore 2013). Through deliberation, the group coheres toward a shared understanding within and integrates a more complete set of relevant considerations, which in turn helps form and revise judgments about the issue in question (Niemeyer et al. 2023). Deliberation thus invigorates the public capacity for thoughtful self-rule under conditions where they can really think to give expression to a meaningful public will.

While deliberation can'happen' anywhere, the conditions for it are likely better when deliberation is intentionally organized and based on a set of procedures which enable citizens to feel that their views are equally and fairly considered. Deliberative mini-publics institutionalize such procedures. These events are designed with the aim of lettting participants follow carefully crafted procedures intended to facilitate balanced information processing and small-group discussions. Numerous studies on the theory of democratic deliberation suggest that participating in deliberative mini-publics leads to increased knowledge about the discussed topic and more thoughtful opinions (cf., Hansen (2004), Andersen and Hansen (2007), Cohen (1997), Elster and Przeworski (1998), and Gutmann and Thompson (2009)). This finding is supported by empirical research as well (cf., Gastil and Dillard (1999) and Barabas (2004)).

Deliberative mini-publics come in many forms and shapes. This project will employ a particular design for deliberative public consultation: Deliberative Polling (for an overview see Fishkin 2019). It is worth discussing how this design differs from other approaches to explain why we chose this design.

2.4 Why deliberative polling?

The main rationale behind the Deliberative Poll is that after the event, participants can see the consequences for valued goals more clearly and weight them more carefully (Luskin et al. 2002). Post-deliberation, citizens' opinions are, at least for some participants, more considered. Participants in a Deliberative Poll receive briefing materials about the deliberation topic, engage in small-group discussions, and ask questions to experts. Taken together, all these individual treatments-within-the-treatment make up the necessary conditions for deliberative democracy to thrive. By participating in steps in the deliberation process, from the pre-survey, reading of the briefing material, discussion with peers, and posing questions to experts on the topics, participants typically learn a great deal about the topic discussed.

Over the last 3 decades, there has been a vast flourishing of deliberative democracy applications employing different designs in countries around the world. These designs include Citizens Juries, Citizens Assemblies, Consensus Conferences and Deliberative Polls (for an overview see Dryzek et al. 2019). At the same time, a considerable critical literature has developed, much of it based on the jury literature. One of the main criticisms has been that the less advantaged will be systematically disadvantaged by a process that appears to privilege those who can best express and argue for their positions (Sanders 1997; Young 2000; Lupia and Norton 2017). To the extent this is the case, the danger is that the results of the deliberation may reflect the power of the more powerful rather than the merits of the arguments being considered. The critique, applied to deliberation in general, is that the more advantaged can be expected to use their better social positions and their greater mastery of the very language and tools of deliberation, to effectively impose their views on everyone else.

These designs vary in terms of sample size, methods of recruitment, methods of conducting the deliberations and of gathering the opinions at the end of the process. Many of the processes mentioned are explicitly consensus seeking. The Citizens Jury is like a jury reaching consensus, but often the product is consensus about more than a verdict; it is a report or extended document of some sort. The same can be said for the Citizens Assemblies, which vary in many particulars, but which have often been aimed at writing a draft law, or a consensus report (Landemore 2020 and for a critical view, Courant 2021). Some of the processes engage relatively small samples (Citizens Juries of perhaps 20 or so; Citizens Assemblies of 50 but sometimes up to 150). Typically, they collect only demographic data at the start so there is no information about the attitudes on the issue that the participants start with. It is rare for these processes to record and transcribe the small-group discussions, so evaluations depend mostly on limited quantitative data, mostly post-deliberation.

Deliberative Polling differs from these other models. It begins with recruitment of a (stratified) random sample, large enough to be meaningfully evaluated in its representativeness at time of recruitment. A pre/post-questionnaire allows for analyses of the opinion changes (if any) at the individual level, without the social pressure of a consensus seeking process. Often there is a separate control group that answers the same questionnaire over the same period. The deliberations take place in moderated small-group discussions where the dialogue can be taped and transcribed for further analysis.

The Deliberative Poll design does not appear to be vulnerable to the critique that the more advantaged strata impose their views (dominate the deliberations). An early study by Alice Siu of five U.S. Deliberative Polls included an examination of the talking time for each participant in the small groups—comparing men vs women and white vs non-white participants. There were no significant differences in talking time (number of words) and minutes spoken in terms of gender. In terms of race, the non-white participants talked more than the white participants (Siu, 2017, p. 122). With moderators and a balanced agenda, there was no evidence of domination, in terms of the degree of participation.

Siu also looked at the movements on the policy issues in the small groups, examining whether they were in the direction supported by the more advantaged. The five projects had 99 small groups and 1474 participants. Considering the more highly educated, those with higher income and those who were white as the more privileged, Siu concludes “we see no consistent movement toward the more privileged in these Deliberative Polls.” (Siu, 2017 p. 123).

A larger study by Luskin, Sood, Fishkin and Hahn looked at the small-group movements in 21 Deliberative Polls, with 2744 group-issue pairs. The projects were geographically and substantively quite varied (US, UK, Australia, China, Bulgaria and EU-wide). They ranged from energy, crime, health care and housing to US foreign policy and the British monarchy.

These topics in the 21 Deliberative Polls generated policy indices permitting an examination of whether the more advantaged were able to cause movements on those policy dimensions in the directions they favored. As Luskin et al. summarized their findings “No matter what the dimension, fewer than 50% of the group-issue pairs move toward the initial mean attitude of the advantaged, meaning that more than 50% move away from it.” This conclusion applied to education, income, race, and gender. If the groups are moving away from the positions of the more advantaged, then the latter can hardly be said to be imposing their views. Rather there is a lot of evidence that the design fosters a deliberation on the merits available to all the strata (see Fishkin 2019 for an overview).

In sum, a Deliberative Poll constitutes an established set of conditions well suited to help citizens increase their knowledge about the use of AI in the public sector. As such, it is a form of public consultation which is well suited for our topic on AI in the public sector, which is not well known in the general public, but where there is a demand among policymakers and others to get citizen input. Moreover, our probability-based sampling and experimental design provide confidence in the statistical generalizability and internal validity of the results, respectively.

3 A deliberative poll experiment on AI

The Deliberative Poll experiment about AI in the public sector took place in Norway during the summer and fall of 2022, with a total of 207 participants. Below we describe in more detail the context of the event, the design, and the results. There are many steps in this research design so we walk through it as follows: we first discuss our case selection; we then show our general experimental design, where the deliberations about AI serve as treatment; we then walk through how the deliberation was set up and carried out; we then go through important details of sampling, fielding, and measurement; and lastly we present our identification and estimation strategy.

3.1 The case of Norway

Norway is characterized by a large public sector which follows the citizens from cradle to grave. The authorities possess large amounts of data on citizens in the population registry, ranging from everything between employment status, education levels, and health status, to voting turnout, criminal record, and country of origin. These data are used by government agencies both for analytical and operational purposes. AI is to our knowledge not currently used for operations that have direct consequences for individual citizens, but models are being developed and studied with the aim of possible future implementation. In the following, we elaborate on three such potential use cases in the Norwegian context.

The Norwegian Labour and Welfare Administration (NAV) aims to transition citizens from welfare to work. NAV develops models to predict the duration of sick leaves, identifying individuals for follow-up. Those likely to be on sick leave beyond 12 weeks may be invited for dialogue about returning to work.

NAV currently selects candidates for these meetings manually. Although everyone is eligible, practical considerations dictate who benefits most from these meetings. By week 16 of sick leave, NAV evaluates the need for a meeting based on responses from the sick person and their employer, and the supervisor's overall assessment.

The expected sick leave duration critically influences this decision. Short-term sick leaves may not require a meeting, whereas those expected to exceed 26 weeks likely do. NAV has developed models using machine learning to estimate sick leave lengths but have not decided on implementing these models yet.

Another potential AI use case is the allocation of admitted refugee. The Directorate of Integration and Diversity (IMDi) currently assigns refugees in asylums to one of Norway's 356 municipalities. Case managers manually process each case, assessing refugees' needs and selecting suitable municipalities for settlement.

Settlement decisions are based on interviews from the refugee reception center and the local municipality. Key factors include country of origin, family composition, relatives in Norway, education, language skills, work experience, and special needs or health services. Refugees' settlement preferences are usually unknown and not prioritized, except in rare cases involving religious considerations. The distribution of resource-intensive cases is managed to avoid burdening any single municipality.

Public servants aim to place refugees in municipalities ready to receive them and where they have the best chances of economic and social integration. International experiences and studies indicate AI could expedite this process and improve outcomes compared to current methods (see Ferweda, Finseraas, and Christensen (2022) for a thorough discussion on data-driven algorithmic refugee settlement in Norway).

The third AI use case involves determining parole eligibility in the Correctional Service. Normally, parole is considered after two-thirds of a sentence is served. The key assessment is whether there's a risk of the convict reoffending, as outlined in the Execution of Sentences Act. Factors include the convict's behavior during imprisonment, criminal history, and efforts to improve. The decision is made by a case manager, based on reports from prison staff, the convict's behavior, and sometimes psychological evaluations. This subjective, labor-intensive process is crucial yet demanding. While Norway has not adopted AI/ML for parole decisions, other countries, like the U.S., use recidivism prediction tools to estimate the likelihood of committing a new crime. These tools, although efficient, have sparked debates over fairness and potential bias against certain inmate groups (Dressel and Farid 2018; Chouldechova 2017; Washington 2018).

All three use cases have direct impact on individual citizens' lives, albeit of varying magnitudes. Parole decisions and refugee assignments in our view clearly fall under what the EU in its developing EU AI Act classifies as High-risk AI systems (Madiega 2023), while a sick leave dialogue meeting arguably has less of an impact on citizens.

After the deliberative event took place, interest in AI has exploded. With this renewed attention, general knowledge among the public has increased (see Supplementary Material, Section B). However, very few citizens in Norway have in-depth knowledge about the particular cases and procedures present above, or about the use of AI in the public sector at large, for that matter (Arnesen and Johannesson 2022). Thus, to study this phenomenon, we must ensure that the general population gains knowledge about the pros and cons of using AI in these three public sector use cases. For this purpose, we organized the Deliberative Poll.

3.2 Experimental design

The setup followed well-established procedures of Deliberative Polls. Here, we would like to highlight a key design deviation: our control group was not a separately recruited survey sample, but rather another group of participants who took part in a Deliberative Poll during the same weekend. During the recruitment process, citizens in our probability-based sample were randomly allocated to being invited either to the topic of AI or the topic of capture and storage of CO_2. Thus, both the treatment and control groups had agreed to take part in a deliberative event the same weekend. Ex ante, the treatment and control groups should, therefore, be more like each other in terms of their motivations, social markers, and exposure to deliberative events than would be the case if the control group consisted of participants who merely took the survey. There are still certain limitations to identification that we discuss later.

Among previous Deliberative Polls, the most similar experimental design is that of Farrar et al. (2010), who split participants in half and randomized the order of topics to be discussed. Surveying the participants in between the topic discussions, the authors were able to measure and compare the attitudinal changes among participants who had deliberated on a set of topics and those who had not. An advantage with this randomized field experimental design is that it cleanly isolates the learning effects of the Deliberative Polling experience itself, assuring that any before-and-after changes stems from the deliberative experience on the topic. Another relevant study was that of Sandefur et al. (2022),

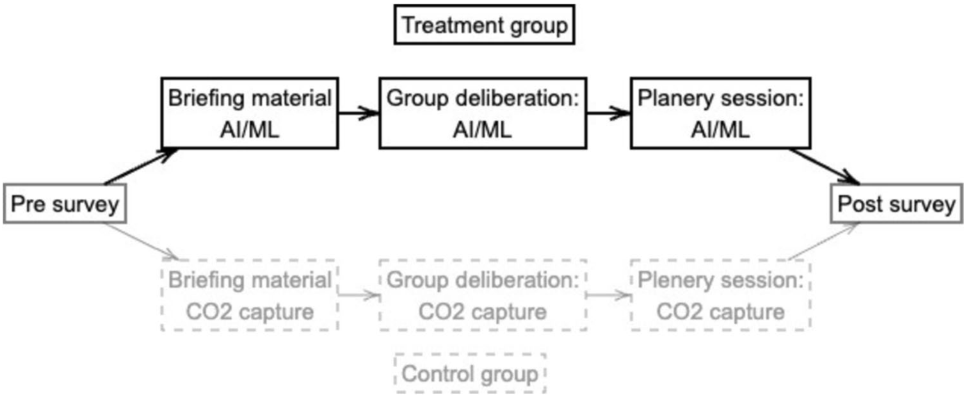
Treatment group
Briefing material AI/ML
Group deliberation: AI/ML
Planery session: AI/ML
Pre survey
Post survey
Briefing material CO2 capture
Group deliberation: CO2 capture
Plenery session: CO2 capture
Control group

Fig. 1 Experimental design

where in addition to a pure control group some participants received information materials only, while others received information materials and took part in deliberation. Only the condition that combined information *and* deliberation produced significant results.

Our design builds on these designs. In our case, the control group for the AI question items never discussed AI, nor received the briefing materials on this topic. They merely responded to question items on AI before and after the deliberative event. These items were embedded in the pre and post-surveys that were part of their deliberation event on the topic of carbon capture and storage, taking place during the same weekends. Vice versa, the AI participants answered questions on both AI and carbon capture, but only deliberated on AI. Thus, the participants served as control groups for each other. Given our focus here on AI, we characterize the participants that discussed AI implementation as our treatment group and the participants that discussed storage and capture of CO_2 as our control group.

This leaves us with only selection effects related to the different topics. Both the treatment and control group answered the same survey both before and after their respective events (Fig. 1).

Since there will be differences in selection effects based on the differences in topic—a deliberation about AI will attract different people, e.g., regarding knowledge about AI, than one about CO2 capture—we cannot identify the treatment effect by comparing the control and treatment group in the post-survey directly without bias. In addition, estimating the effect by comparing pre- and post for the treatment group can also be biased because answering the pre-survey may itself increase the knowledge about AI even without any deliberation. To identify the causal effect of the deliberation, we use difference-in-difference estimator: we compare the change from pre- to post-survey between the treatment and control group. We are thus only making the critical but reasonable assumption of parallel trends. That is, we assume that the treatment group would have changed similarly to the control group (e.g., with regards to their knowledge about AI) had they instead participated in the control deliberation. This allows us to identify the Average Treatment Effect on the Treated (ATT), i.e., the causal effect of the deliberation for the participants in the treatment group.

3.3 Measurement

To measure participants' preferences when it comes to the use of AI/ML in public administration, we field three survey items. We ask whether participants think, based on their knowledge, whether the government should open for the use of AI/ML in three areas where the use of AI-based decisions is possible in the future. These are refugee settlement, sick leave dialogue meeting with NAV, and parole. Participants were asked to respond on a bipolar seven-point scale, ranging from "support very strongly" to "oppose very strongly". In addition, we also included a "do not know" option. Post treatment, we calculate the average treatment effect of the treated (ATT) for these individual survey items. This tells us whether there are different changes (if any) in opinions among the treated units.[3]

To check that our Deliberative Poll had similar knowledge effects observed in previous deliberative events, we ask participants about their knowledge about AI/ML, as well as their knowledge about the current procedures of the three cases. Participants are asked to respond on a unipolar five-point scale, ranging from very good knowledge to no knowledge at all. We also include a "do not know" option. The choice of using self-reported knowledge as our knowledge measure is based on the aim to keep attrition at a minimum and the consideration that factual knowledge questions in online surveys tend to suffer from respondent dishonesty (Höhne et al. 2020; Rapeli 2022).

[3] See Supplementary Material, section E.

3.4 Execution: sampling, fielding, and measurement

Participants were recruited in two waves—one in June 2022 and one in September 2022—resulting in 207 participants who completed the entire process. Using the national population registry, we randomly drew a sample of Norwegian citizens. The population registry constitutes a near perfect sampling frame, thus fulfilling the important criterion of equal inclusion of citizens. What is more, in contrast to non-probability samples, the quality of probability-based samples can be assessed using an established framework which makes it possible to compute the accuracy of estimates and gives a universal validity to the estimation (Cornesse et al. 2020; Groves and Lyberg 2010).

We recruited participants into two parallel sessions where participants were asked to discuss different topics: one related to capturing and storing CO_2 and one related to the use of artificial intelligence in the public sector. In line with recommendations on setting the right preconditions for online deliberation (Strandberg and Grönlund 2012) invitees were offered a cash incentive upon completed participation (NOK 1000,-/USD ~100). The recruitment resulted in total in 207 participants, whereby 118 deliberated on AI and 89 deliberated on CO_2 storage and capture.[4]

Given our access to population registry data, we can compare the social background characteristics of the participants with that of the gross sample. Since the gross sample is large and randomly drawn from the entire population registry, we assume that it very accurately reflects the target population of residents in Norway. Comparing the social background characteristics of the participants with the gross sample, we observe that they were quite representative in terms of age, gender, income, and whether they were born in Norway or outside. The share of participants with higher education and the share who voted in the 2021 national parliamentary election were significantly higher than in the gross sample.[5]

The fielding of the pre- and post-surveys was done in-house using Qualtrics. The pre-survey was fielded a week before the day of deliberation. After the survey was completed, participants were given briefing material in the form of a 25-page document about the topic discussed. The AI group was given information about AI/ML and the other was given information about the capture and storage of CO_2 (Fig. 2).

Citizens who agreed to participate would first receive an online pre-survey where they were asked to respond to a 15-min questionnaire related to the deliberation topic as well as some questions about their background. After the deadline for responding had passed, all respondents who had completed the survey were provided with briefing materials about the topics they were going to discuss. In the AI briefing material participants were given some general information about what AI is and what it can be used for. They also received background information on the three use cases where AI could be used in the future, described in Section 3.1. Participants were shown several policy proposals and statements with pro and con arguments related to each use case, to be described in more detail under Section 4.2.

On the day of the deliberation, participants logged into an online platform—The Stanford Online Deliberation Platform (SODP). This web-based platform facilitates constructive discussion with the use of an automated moderator, stimulating participants in video-based discussions to consider arguments from both sides of all proposals, maintaining equitable participation and civility in the discussion. On the platform they were presented with the schedule for the day and a video wrapping up the content in the briefing materials.

Participants were assigned to discussion groups, ranging from 6 to 12 in size, and asked to discuss the same policy proposals and statements described in the briefing material. There were two group deliberation sessions: one in the morning, and one in the afternoon. The morning session topics were revolved around the welfare-to-work and refugee use cases, where the participants discussed the following proposals: (1) AI should be used to decide settlement of refugees, (2) AI should be used to decide sick leave, (3) refugee settlements should be distributed by work region not municipalities, (4) AI technology development for refugee settlement should reflect all citizens, and (5) all available information should be used to predict sick leave. In the afternoon session, the participants discussed the pros and cons of the proposals that (6) AI should be used with humans having the final say, (7) accuracy of AI should be judged against the current human models, (8) we should use AI to decide parole, (9) the share of black and white inmates classified as high risk with respect to recidivism should be the same, and (10) we should be as accurate as possible when giving someone parole. After about an hour of deliberation in each session, the groups were asked to write down questions resulting from their deliberation that were to be asked to several experts on AI or

[4] The study was conducted in two waves because the number of participants did not meet our sample size target. 118 citizens participated in the first wave (50 in the CO_2 storage and capture group and 68 in the AI/ML group), and 89 participated in the second wave (39 in the CO_2 storage and capture group and 50 in the AI/ML group). In the analysis, we pool the participants of the two waves into one. Robustness tests show that there are no significant differences in the results between the participants in the two waves, see S.I for more details.

[5] See Supplementary Material, section C for details.

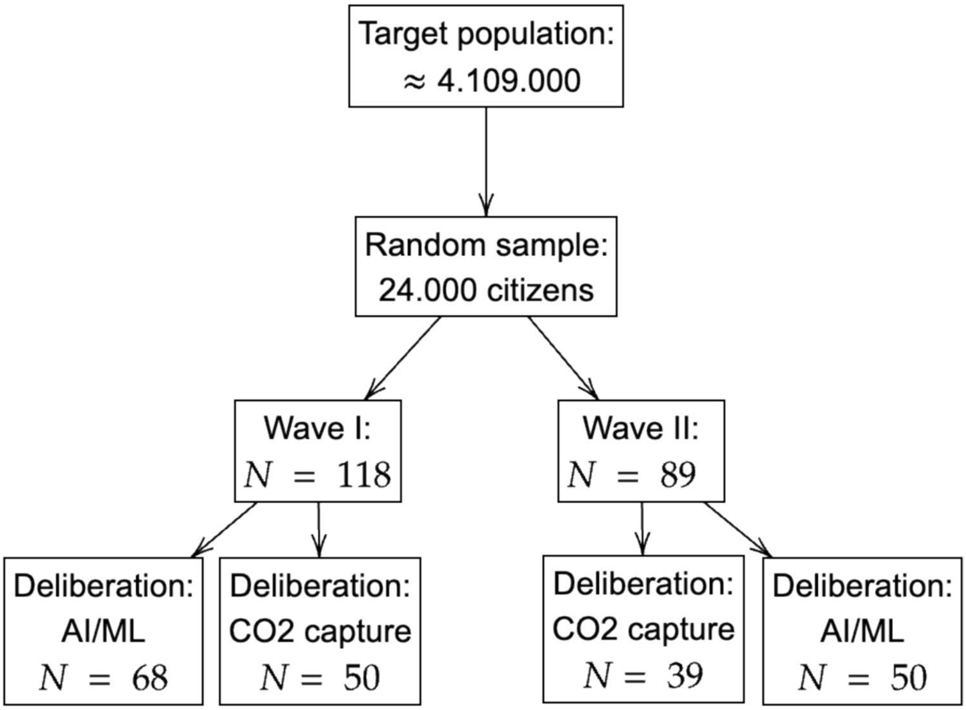

Fig. 2 Sampling and group assignment

the cases discussed. Each group prioritized two questions to be asked during the plenary session.

In the plenary session, the experts answered these questions and, on occasion, questions from the participants posted in a live chat. The experts were recruited with the aim of providing balanced feedback on the pros and cons of using AI in the circumstances, as well as providing factual knowledge on AI and on the use cases. The experts had either been involved in research on AI or worked in the relevant government institutions. After the plenary session, participants were randomly assigned to new discussion groups and continued to discuss new policy proposals (proposals 6–10). After this second group session, a new round of questions was posed to experts in a new plenary session, which concluded the deliberation. When the plenary session ended, participants were asked to take the exact same online survey again. In all, the deliberative event lasted for five hours, including breaks.

3.5 Identification and estimation strategy

After all participants completed the pre-survey, deliberation, and the post-survey, we measured the average treatment effect of the treated (ATT) of a change in their response to a specific survey question using a difference-in-difference design. Formally,

$$ATT = E\left[y_{i1} - y_{i0}\right] \tag{1}$$

where ATT is the average treatment effect of the treated, y is the difference for i respondent in the pre- and post-deliberation survey. We estimate the treatment effect with a difference-in-difference (DiD) estimator. Formally,

$$y_{it} = \gamma_{s(i)} + \lambda_t + \delta I(...) + \in_{it} \tag{2}$$

where $y_i t$ is the dependent variable for each respondent i at time t, $s(i)$ denotes the treatment or control group respondent i belongs to, and δ is our quantity of interest, i.e., the treatment effect, and $I(...)$. denotes the dichotomous treatment and time variables.

4 Results

We show in this section the experimental results which establish that participants became more positive towards the government's use of AI/ML when they provide services related to NAV's dialogue meeting, refugee settlement, and granting parole (supporting H1). After presenting the experimental analysis, we qualitatively analyze the transcripts of the group session recordings where the participants deliberate on specific proposals related to the use cases.

4.1 Experimental treatment effects

Result I The deliberation made participants more positive towards using AI/ML in the welfare-to-work case (sick-leave dialogue meeting).

Before the event, 36 percent of participants 'somewhat support' that NAV use AI to decide when to invite people for a dialogue meeting when they are on sick leave, 36 percent18 percent and 14 percent report that they “neither oppose or support” or “oppose somewhat”, respectively ($\mu = 4.315$, $\sigma = 1.465$). Seven participants answer that they “do not know” when asked whether they want the government to open for the use of AI in this area.

After the event, we observe a substantial shift in the responses. The mean value increases with nearly 1 point on the answer scale, indicating increasing support for the use of AI in NAV ($\mu = 5.174$, $\sigma = 1.126$) and three participants report that they “do not know”. 84 percent of the participants answer that they support the use of AI in NAV, post-treatment (47 percent answer “support somewhat” and 37 percent answer “support strongly)”. A graphical presentation of the post-treatment response distributions is provided in Supplementary Material Section B.1. We confirm that this is a result of our intervention—the deliberation event—by inspecting Fig. 3. Here, the average change in opinion is almost one point on the seven-point answer scale ($\delta = 0.854$, $p < 0.05$), which shows that post-deliberation, participants are more positive towards the use of AI in NAV’s services in comparison with the control group.

Result II The deliberation made participants more positive towards the government’s use of AI/ML for refugee settlement tasks.

Pre-treatment, on the question related to settlement of refugees 35 percent of the participants answered that they “support somewhat” the governments use of AI/ML when they decide where refugees should be settled, while 23 percent and 22 percent answered “neither support or oppose” and “oppose somewhat”, respectively ($\mu = 4.27$ and $\sigma = 1.26$). Three participants answered “do not know”. After the event, 45 percent “support somewhat” and 37 percent “support strongly” that the government should use AI/ML to settle refugees ($\mu = 5.22$ and $\sigma = 1.12$). One person reported “do not know” after the deliberation event. Again, we observe a treatment effect of around one point on the seven-point scale ($\delta = 0.938$, $p < 0.05$), showing an increase in support for the use of AI when the government decides where refugees should be settled.

Result III The deliberation made participants more positive towards the use of AI/ML for parole decisions.

For the last question, we see some of the same trends as with the previous survey items in the descriptive results. Albeit the mean value in the pre-treatment responses is somewhat lower ($\mu = 3.583$, $\sigma = 1.589$). Most participants, some 28 percent, “support somewhat” the use of AI/ML when the authorities give parole. Yet, 49 percent either “oppose somewhat”, “oppose strongly”, or “oppose very strongly”. 22 percent “neither oppose or support”. Substantially, we observe a change in the mean value of response on this question ($\mu = 4.4874$, $\sigma = 1.436$) with an increase of 0.9 in the post-treatment survey. Here, 43 percent “support somewhat” and 20 percent “support strongly” that the government open up for the use of AI/ML when deciding who to give parole to. This observational analysis also holds when we apply the more rigorous difference-in-difference design: In Fig. 3, we observe a causal effect of around one point increase ($\delta = 1.002$, $p < 0.05$) in the support for AI when the government decides parole.

Moreover, as expected from previous studies, participants in the treatment group reported to have gained more knowledge on AI and ML, and on the cases under scrutiny (Fig. 4).

Analysis of weighted data yield the same results as shown in Figs. 3 and 4 (see supplementary material C1).

4.2 Text analysis of group deliberations

To get a deeper understanding about how the participants were thinking about the use cases, and the pros and cons of using AI in these contexts, we conducted a qualitative thematic analysis of the group deliberations. Based on anonymized transcripts of 24 group deliberations, we

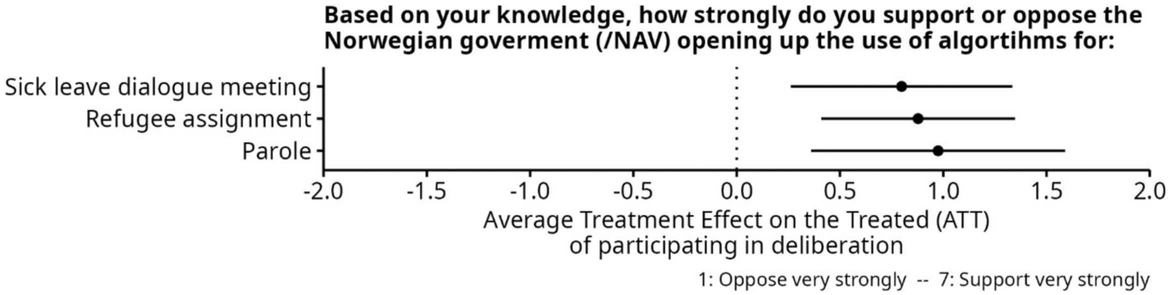

Fig. 3 Support for the government’s use of AI. Note: The figure shows the average treatment effect for the treated units (ATT) after reading the information material, participating in the deliberation, and listening to the expert session. The vertical-axis represents the wording of the question and the horizontal axis shows the average change in participants' answer to a specific question, post-treatment. The horizontal bars show 95% confidence intervals. *N* respondents = 207; *N* observations = 414

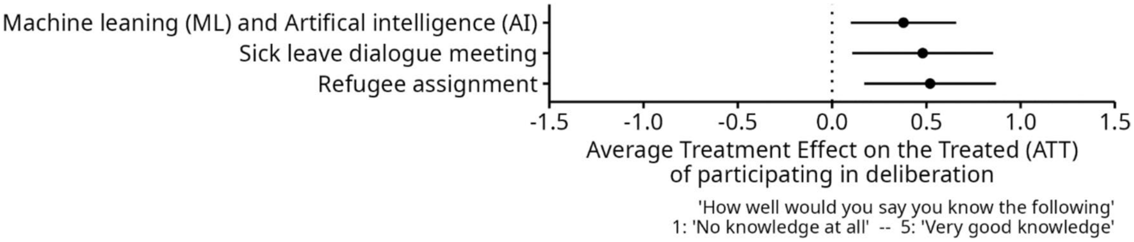

Fig. 4 Participants' knowledge about AI and government services. Note: The figure shows the average treatment effect for the treated units (ATT) after reading the information material, participating in the deliberation, and listening to the expert session. The vertical axis represents the wording of the question and the horizontal axis shows the average change in participants answer to a specific question, post-treatment. The horizontal bars show 95% confidence intervals. *N* respondents = 207; *N* observations = 414

have a text corpus consisting of 96,714 words and 1854 speeches.

The large language model used to analyze the transcript is called the Deliberation Analyzer. The Deliberative Democracy Lab at Stanford University and Fileread created the Deliberation Analyzer, a fit-for-purpose natural language processing pipeline to analyze deliberations using two language models: BERT and GPT-4. At the time of processing data for this paper, BERT-based models remain among the most state-of-the-art for analyzing language. For example, it performs at the highest benchmarks on tasks such as emotional analysis, the task of identifying strong emotions in language. Although OpenAI's GPT family of models has captured the public imagination as a versatile all-purpose language model, it still sometimes falls behind BERT for analysis. However, without a doubt, GPT-4 is still the best generative language model—excelling at tasks such as summarization and storytelling. For maximum efficacy, Fileread uses BERT for deliberation analysis. The analysis is then piped to GPT-4 for interpretation and summarization. Further iterations of the Deliberation Analyzer have used other language models as well. Using this tool, we systematically extracted the pro and con arguments expressed during the discussion for each of the ten proposals. Our thematic analysis was based on both the transcripts, and these systematically extracted arguments. Table 1 displays the proposals along with snippets from the transcripts exemplifying the pro and con arguments made by the participants.

For overview purposes, a count of pro and con arguments for each proposal is shown in Table 2. As the table shows, some of the proposals were directly related towards discussing the use of AI in the three use cases, while other proposals were more generic, albeit still relevant for the use cases in some capacity. We see from the summary statistics that arguments against using AI are actually more frequent than arguments in favor. This serves as a reminder that the quality of arguments weigh more heavily than the quantity. We also note that fully automated AI systems were frequently argued against during the group deliberations.

Our thematic analysis of the transcripts reveals that while the discussions and specific arguments varied between the different proposals, there were several reoccurring themes. These themes often revolved around trade-offs between different considerations related to using AI in the public sector. One major recurring theme is the balance between AI's (potential) efficiency and the ethical concerns around reducing human involvement. In general, there was a strong emphasis on the importance of having humans in the loop, in conjunction with AI, as opposed to fully automated decision-making. Another recurring theme is the balance between using more data for better decisions and ethical concerns about privacy, biases, and data security. Such ethical concerns are prominent, particularly regarding the potential for bias in AI systems. Many expressed that AI might perpetuate or even exacerbate existing social biases, especially in sensitive areas like refugee settlement and healthcare. Data privacy was also a strong theme, with concerns about how much personal data is being used, who has access to it, and the potential for misuse. However, it was also argued that with proper safeguards and regulations, AI systems can be designed to protect privacy while still being effective. In general, while there was significant support expressed for the use of AI in public sector decisions, concerns about ethical implications, data integrity, and the role of human judgment are key issues that were discussed.

Exactly what the participants learned in the Deliberative Poll that made them change their minds cannot be identified. However, the transcriptions of the group discussions yield some insights. This quote from one of the participants sums up well both the average position of the participants, as well as a qualifying condition that was reiterated several times across the groups during the small-group discussions: "Yes, I am partly positive for the use of algorithms and data processing. It can make the process better as long as it is done properly. But I am then skeptical that it should only be data processing, there must be a human factor inside it." As

Table 1 Proposals discussed and example quotes from transcripts arguing for and against the proposal

Proposal	Example against	Example for
Accuracy of AI should be judged against the current human models	Algorithms can make decisions much faster and actually better than humans. There is perhaps a tendency to overestimate people's ability to make decisions	If the algorithms are good enough, then that is where we're headed, implying that AI accuracy should be judged against current human models to determine if they are good enough to replace humans. In the initial phase, you have to include humans, and then assess how well these algorithms do in comparison, further suggesting that AI accuracy should be evaluated against human performance. If algorithms perform well enough compared to humans, then we can replace more and more tasks with algorithms, indicating that the benchmark for AI adoption should be current human-level performance
AI should be used to decide settlement of refugees	The data used in an algorithm are far from exhaustive and captures very little of the person, such as only age and professional background	It will be much quicker to distribute refugees if you use machine-based assistance
AI should be used to decide sick leave	It is important to consider how much data NAV has access to and who has access to that data, because it often contains very personal information	If algorithms and machine learning can give individual NAV employees more time to follow-up on their clients, that is a good thing
AI should be used with humans having the final say	Algorithms are a good thing because they take away human error. The algorithm should not be racist	The democratic basis must be present in determining how the algorithm should be, implying human oversight and decision-making. Experts who know this field professionally should take the guidelines further based on a basic definition of desired algorithms, suggesting that humans should have the final say in implementing and adjusting the AI system
AI technology development for refugee resettlement should reflect all citizens	It should be people who know this professionally. It is no good us as amateurs in this field sitting around laying down guidelines that do not work well. Therefore, we need to have a mix of democratically representative selections of interested citizens and experts	There has to be some kind of steering group that has control over how the algorithm works. The technical part of coding is one side, but what is important is that there is a steering group that has control over how it works. This argument suggests that a diverse steering group representing the citizens should oversee and control the AI development, even if they are not involved in the technical coding itself
All available information should be used to predict sick leave	It is important to consider how much data NAV has access to and who has access to that data, because the data may contain very personal information One should be critical of what information is shared, presumably because some personal information should be kept private	They should have as much information as possible to make it accurate, because we want it to be as accurate as possible
Refugee resettlements should be distributed by work region not municipalities	There is a number of guidelines and constraints that make political control essentially impossible when using algorithms and work regions for refugee resettlement, and political control is needed	The algorithms should take into account experience and where the newcomers have the best chance of success, and distribute refugees accordingly. If the algorithms are more accurate for regions than for municipalities, then refugee resettlement should probably be distributed by region
The share of black and white inmates should be the same	Ideally, the algorithm should not adjust differently based on race, such as "okay you're white so we'll adjust a bit like this, you're black, so we'll adjust a bit like this"	Statistics show it is harder for non-white people to get parole, which suggests parole should be granted more equally

Table 1 (continued)

Proposal	Example against	Example for
We should be as accurate as possible when giving someone parole	It is very important that someone who is dangerous is not released into society, it is a greater risk for the majority than that a person who should have been granted parole has to do time	It is completely crazy to give one group more wrong decisions so that the number of wrong decisions will be the same between several demographic groups
We should use AI to decide parole	Using algorithms to decide parole means you do not see the whole person. Not seeing the whole person when deciding parole can be difficult at times	It is good to have data to relate to when making parole decisions. It can be difficult for a person to be completely objective based on the sentences that individuals have. A semi-automatic process using AI can be a good solution to get a more objective assessment

became clear during the Deliberative Poll, none of the three use cases was intended to be fully automated AI decision processes. This may have been comforting for the participants and helped push them in the direction of being more supportive of using AI in these cases. There was furthermore the acknowledgment during the deliberative process that the competition of the AI models are not perfection, but rather the imperfect, flawed human based decisions of the existing system: Human administrators are limited in their capacities and flawed in their pursuit of efficiency, effectiveness, and equitable administration of the law (Bullock 2019). Against this benchmark, supporting the use of AI becomes easier than when the benchmark is a utopian, perfect decision-making system.

5 Discussion and conclusion

The results show that participants became more positive towards the use of AI in the cases they discussed. In addition, the effect sizes are quite substantial; between 0.85 and 1.002 on a seven-point scale, meaning that, on average, respondents became nearly one point more positive compared to the control group. Despite the rather low sample size, the effects are all clearly statistically significant. Furthermore, participants also reported that they gained more knowledge about AI/ML and government services after participating in the event, supporting our expectations from the existing literature.

What was until now a correlational relationship has been strengthened with our experimental findings pointing in the same direction: The more people learn about artificial intelligence and machine learning, the more they support it, even in decision tasks that are quite consequential for individuals in society. The shift in attitude was significant in comparison with that of the control group, which also took part in a deliberative event on a non-AI topic. The samples size restricts us from meaningfully exploring heterogeneous treatment effects among participants.

Whereas the participants became more positive towards AI in our cases, we are nonetheless wary of making strong claims concerning their external validity. AI can be many things, varying in terms of the risks the decisions imply for citizens, what role AI plays in the decision-making process, and how well they perform, to name just a few considerations. As for internal validity, we have made considerable efforts in providing balanced briefing materials, recruiting experts with a diverse background, using neutral language in the questionnaire, etcetera. In spite of this, we recognize that we as organizers have an agenda-setting role that has the potential to influence the views of the participants. Thus, for the sake of transparency, all transcriptions from recordings of the small-group discussions and expert plenary sessions

Table 2 Number of different pro and con arguments during group discussions for each proposal

Proposal	Against (%)	For (%)	Diff (%)
Accuracy of AI should be judged against the current human models	49	51	+2
AI should be used to decide resettlement of refugees	56	44	– 12
AI should be used to decide sick leave	56	44	– 12
AI should be used with humans having the final say	14	86	+72
AI technology development for refugee resettlement should reflect all citizens	33	67	+34
All available information should be used to predict sick leave	63	37	– 26
Refugee resettlements should be distributed by work region not municipalities	56	33	– 12
The share of black and white inmates should be the same	69	31	– 38
We should be as accurate as possible when giving someone parole	28	72	+44
We should use AI to decide parole	63	37	– 26

are available for anyone who would like to study how the deliberative event unfolded, as well as the original briefing materials the participants read and the questionnaire they answered (see the online Supplementary Information).

At least two of the three use cases under scrutiny will likely be categorized as high-risk cases under the EU AI Act proposed regulation, respectively filing under the topics of Law enforcement, and Migration, asylum, and border control management. One implication from this study is for policymakers to think about establishing procedures which allow for democratic inputs before implementing AI tools in the public sector when they have direct influence on individual citizens' lives and are considered high-risk use cases. It is in our view part of a democratic process where citizens are awarded with the capacity to form considered opinions about important topics that regulate their lives. There is much to gain for governmental authorities by informing and involving citizens prior to introducing AI in their decision processes. Citizens may initially approach AI in the public sector with a sound skepticism, yet they are open to changing their minds when they learn more about what these tools do and how they will be implemented into existing bureaucratic procedures. As students of public opinion, we are agnostic about what way the participants' attitudes will lean after a successful deliberation, but a better-informed citizenry is always a good thing.

5.1 Limitations

One of the limitations of the Deliberative Polling design is also one of its merits. It requires that the deliberators constitute a (stratified) random sample representative of the population. The idea is to represent what the people would think, if they were to consider the issue under stipulated good conditions (balanced briefing materials, moderated small-group discussions, plenary sessions with competing experts to get their questions answered, absence of social pressure to agree with everyone else, gathering of conclusions in confidential questionnaires). However, the need for a representative sample means that populations who are not large enough to be represented are not included (or just barely included). Does it undermine the results if they are not included—particularly if the topic of discussion directly concerns the population in question? Kim et al. (2018) studied two national Deliberative Polls where the subject was policies toward a vulnerable population (Aboriginals in Australia and the Roma in Bulgaria). They examined the opinion changes in small groups that included members of the vulnerable population compared to those without (group assignments were random). The basic finding was twofold. All the small groups in both projects moved in the direction of policies that would favor the minority group. But the groups with the vulnerable population included moved somewhat more in that direction. Therefore, there is an effect but it does not yield notably different conclusions about the policies than those resulting from the other groups without members from the vulnerable group (Kim et al. 2018).

Since this project aimed at policy relevance based on the considered judgments of a representative sample of the national electorate, it proceeded without additional sampling for refugees or parolees (the two vulnerable populations potentially affected by the topics). We support the argument that such affected groups often should be given special consideration before AI tool are implemented. Whereas our aim with the presented Deliberative Poll was to gauge the perspectives of the general population, other designs for deliberative experiments could be designed to shed light on the effects of inclusion of these groups in the deliberations. The target populations would be different from ours (national electorate vs. inmates or refugees), and the recruitment strategy for achieving representative samples of those target populations would be different from the one we applied. Although beyond the scope of our study, we encourage such initiatives and see them as important, complementary studies to ours.

Another critique of deliberative democracy, distinguishable from the domination argument, is that deliberative democracy methods aim for rational consensus devoid of the emotions that drive politics. Chantalle Mouffe has argued that politics is inherently agonistic and deliberative democracy is inappropriate because it pushes for consensus that covers over the essential conflicts (Mouffe 2005), chapter four). First, note that while this may be true for the consensus seeking designs discussed earlier (Citizens Juries, Citizens Assemblies) it is not true of Deliberative Polling. This design does not seek consensus but collects the conclusions in confidential questionnaires. If a consensus emerges it will be apparent in the post-deliberation data. But if there is an intractable division that will also be apparent in the data. Furthermore, deliberation on this design does not neglect emotions. Most obviously, empathy for vulnerable populations drove the pro-Aboriginal and pro-Roma policy movements in the projects cited above (Kim et al. 2018; Fishkin 2019).

The choice of using self-reported knowledge as our knowledge measure is based on the aim to keep attrition at a minimum and the consideration that factual knowledge questions in online surveys tend to suffer from respondent dishonesty (Höhne et al. 2020; Rapeli 2022). A drawback with the subjective knowledge measure, though, is that we leave up to each participant to assess their knowledge about the topic. The accuracy with which participants can assess their knowledge on the topic is likely to vary between participants. That said, our measure of interest is the within-subject variation before and after the deliberative event, where we only must assume that each individual participant is consistent in the way they assess their own knowledge. We cannot fully rule out experimenter demand effects, i.e., that the participants report higher knowledge on the topic because that is what they believe we as organizers want to see. The participants were, however, never informed that we expected them to gain knowledge about the topic. Rather, we argue knowledge gains is a natural consequence of spending time reading about, deliberating, and listening to experts on the topic. It is highly likely that anyone who follows such a course of action will become more knowledgeable about a topic, irrespective of how it is measured. As expected, the self-reported knowledge measure confirms this.

5.2 Contributions and further research

AI is quickly becoming a readily available tool with enormous potential for making a more efficient public sector. Normative discussions about the fairness, accountability, and transparency of AI-based decision-making are abound, yet little is known about citizens' considered views. The AI development particularly in the public sector needs to be socially aligned, and our contribution to the demand for empirical research on specific AI use cases in the public sector is a Deliberative Poll on using AI for refugee reallocation, parole decisions, and eligibility for welfare programs. This deliberative mini public has given citizen inputs to AI with an emphasis on the democratic ideals of deliberation, participation, and political equality. The aim has been to increase the competence among a representative sample of the population to enhance their capacity to consider normative, technical, and political implications of using AI in governance. Throughout a full day, participants ran through an agenda touching upon several aspects of using AI in the three potential use cases. In their own eyes, the deliberative event made them both more knowledgeable about AI/ML, and more supportive of its use in three potential use cases.

The main contributions of this study are hence (1) that it introduces a framework for increasing public knowledge and involvement in AI policymaking in the public sector through Deliberative Polling, which includes education on AI/ML, peer discussions, and expert interactions. (2) The study provides experimental evidence, using a sample of 207 participants, that increasing knowledge about AI leads to greater public support for its use in government decision-making. It explores attitudes toward AI in three specific public sector use cases: refugee settlement, welfare-to-work programs, and parole decisions, finding that informed citizens are more supportive of AI integration in these areas.

We suggest several avenues for further research. We encourage similar Deliberative Polls to be conducted in diverse geographical regions and cultural contexts to assess whether the findings are consistent across different populations, particularly in countries with different levels of AI adoption and public sector structures. Expanding research to include a wider variety of AI applications in the public sector, beyond refugee reallocation, welfare-to-work programs, and parole would also be useful to investigate the generalizability of the results. This could include areas such as healthcare, education, or policing, to see if public attitudes vary depending on the specific application. Future studies could include more targeted deliberative processes involving vulnerable populations directly affected by AI decisions (e.g., refugees or parolees) to understand how their perspectives might differ from those of the general population. These research avenues would help deepen the understanding of public attitudes toward AI in governance and improve the design and implementation of AI policies in the public sector.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00146-024-02104-w.

Funding Open access funding provided by NORCE Norwegian Research Centre AS. Norges Forskningsråd, 314114, Sveinung Arnesen.

Data availability The survey data and a replication file are publicly available in the Harvard Dataverse repository, https://doi.org/10.7910/DVN/98RBQV.

The group discussions and expert session transcripts are available upon request to the authors.

Declarations

Conflict of interest On behalf of all the authors, the corresponding author states that there is no conflict of interest.

References

Alice S (2017) Deliberation & the challenge of inequality. Daedalus 146(3):119–128

Andersen VN, Hansen KM (2007) How deliberation makes better citizens: the danish deliberative poll on the euro. Eur J Polit Res 46(4):531–556

Araujo T, Helberger N, Kruikemeier S, De Vreese CH (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & Soc 35(3):611–623

Arnesen, Sveinung, and Mikael Poul Johannesson (2022) Demokratiske algoritmer. NORCE Report 18–2022. https://norceresearch.brage.unit.no/norceresearch-xmlui/bitstream/handle/11250/2995353/norce+helse+og+samfunn%2C+rapport+18-2022.pdf?sequence=1. Accessed 23 Oct 2024.

Andre B, John SD, Mansbridge J, Mark EW (2018) The Oxford handbook of deliberative democracy. Oxford University Press

Anneke Z, Chen Y-C, Salem F (2021) Implications of the use of artificial intelligence in public governance: a systematic literature review and a research agenda. Government Inform Quart 38(3):101577. https://doi.org/10.1016/j.giq.2021.101577

Barabas J (2004) How deliberation affects policy opinions. Am Political Sci Rev 98(4):687–701

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif l Rev 104:671

Beckman L, Rosenberg JH, Jebari K (2022) Artificial intelligence and democratic legitimacy. The problem of publicity in public authority. AI Soc 39:1–10

Binns R (2018) Algorithmic accountability and public reason. Philosophy Technol 31(4):543–556

Bullock JB (2019) Artificial intelligence, discretion, and bureaucracy. Am Rev Public Admin 49(7):751–761

Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5(2):153–163. https://doi.org/10.1089/big.2016.0047

Cohen, Joshua. 1997. "Procedure and substance in deliberative democracy." Deliberative democracy: Essays on reason and politics, 407

Cornesse C, Annelies GB, Dutwin D, Jon AK, Edith DDL, Legleye S, Pasek J, Pennay D, Phillips B, Joseph WS et al (2020) A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. J Surv Stat Methodol 8(1):4–36

Courant D, 2021 The Promises and disappointments of the french citizens' convention for climate. Deliberative Democracy Digest

Christopher S, Lunich M (2020) Artificial intelligence for political decision-making in the European Union: effects on citizens' perceptions of input, throughput, and output legitimacy. Data Amp; Policy 2:16. https://doi.org/10.1017/dap.2020.19

de Fine LK, Jenny DFL (2020) Artificial intelligence, transparency, and public decision-making. AI Soc 35(4):917–926

Dickson ES, Gordon SC, Huber GA (2015) Institutional sources of legitimate authority: An experimental investigation. Am J Political Sci 59(1):109–127

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4(1):eaao5580. https://doi.org/10.1126/sciadv.aao5580

Dryzek JS, Bächtiger A, Chambers S, Cohen J, Druckman JN, Felicetti A, Fishkin JS, Farrell DM, Fung A, Gutmann A, Landemore H, Mansbridge J, Marien S, Neblo MA, Niemeyer S, Setälä M, Slothuus R, Suiter J, Thompson D, Warren ME (2019) The crisis of democracy and the science of deliberation. Science 363(6432):1144–1146

Elster J, Przeworski A (1998) Deliberative democracy, vol 1. Cambridge University Press

David E (2009) Democratic authority. Democratic authority. Princeton University Press

Farrar C, Fishkin JS, Green DP, List C, Luskin RC, Paluck EL (2010) Disaggregating deliberation's effects: an experiment within a deliberative poll. Br J Political Sci 40(2):333–347

Ferweda Jeremy, Henning Finseraas, and Dag Arne Christensen 2022 "The feasibility of using data-driven algorithmic recommendations for refugee placement in Norway"

Fishkin JS (2019) Democracy when the people are thinking. Oxford University Press, Oxford

Fuerth L (2011) Operationalizing anticipatory governance. Prism 2(4):31–46

Gastil J, Dillard JP (1999) Increasing political sophistication through public deliberation. Political Commun 16(1):3–23

Grgic-Hlaca, Nina, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. "Human perceptions of fairness in algorithmic decision making: a case study of criminal risk prediction." In: Proceedings of the 2018 world wide web conference, 903–912

Groves RM, Lyberg L (2010) Total survey error: past, present, and future. Public Opin Q 74(5):849–879

Grzymek, Viktoria, and Michael Puntschuh. 2019. "Was Europa u¨ber Algorithmen weiss und denkt"

Guston DH (2014) Understanding 'anticipatory governance.' Soc Stud Sci 44(2):218–242

Gutmann A, Thompson DF (2009) Democracy and disagreement. Harvard University Press

Hansen KM (2004) Deliberative democracy and opinion formation. University Press of Southern Denmark Odense

Karem HJ, Cornesse C, Schlosser S, Couper MP, Blom AG (2020) Looking up answers to political knowledge questions in web surveys. Public Opinion Quart 84(4):986–999

Kantar 2023 How do people feel about AI. Technical report. https://github.com/AdaLovelaceInstitute/how-do-people-feel-about-ai

Kim N, Fishkin JS, Luskin RC (2018) Intergroup contact in deliberative contexts: evidence from deliberative polls. J Commun 68(6):1029–1051. https://doi.org/10.1093/joc/jqy056

Kim S, Gronlund K (2012) Online deliberation and its outcome—evidence from the virtual polity experiment. J Inform Technol Politics 9(2):167–184

Konig PD (2022) Citizen conceptions of democracy and support for artificial intelligence in government and politics. Euro J Political Res 62:1280–1300

Kroll Joshua Alexander 2015 "Accountable algorithms." PhD diss., Princeton University

Landemore H (2013) Deliberation, cognitive diversity, and democratic inclusiveness: an epistemic argument for the random selection of representatives. Synthese 190(7):1209–1231

Landemore H (2020) Open democracy: *reinventing popular rule for the twenty-first century*. Princeton University Press, Princeton

Lee, Min Kyung, and Su Baykal. 2017. "Algorithmic mediation in group decisions: fairness perceptions of algorithmically mediated vs. discussion-based social division." In: Proceedings of the 2017 acm conference on computer supported cooperative work and social computing, 1035–1048

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: people prefer algorithmic to human judgment. Organ Behav Hum Decis Process 151:90–103

Lupia A, Anne Norton A (2017) Inequality is always in the room language & power in deliberative democracy. Daedalus 146(3):64–76

Luskin RC, Fishkin JS, Jowell R (2002) Considered opinions: deliberative polling in Britain. Br J Political Sci 32(3):455–487

Madiega Tambiama 2023 "Artificial intelligence act." European parliamentary research service, https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRSBRI(2021)698792%20EN.pdf

Mansbridge Jane, James Bohman, Simone Chambers, Thomas Christiano, Archon Fung, John Parkinson, Dennis F Thompson, Mark E Warren (2012) A systemic approach to deliberative democracy." in Deliberative systems: Deliberative democracy at the large scale, Volume 1, p. 1–26.

Mouffe C (2005) The democratic paradox. Verso, London

Noordt Tangi L. van, M. Combetto, D. Gattwinkel, F. Pignatelli, and AI Watch. 2022. European landscape on the use of artificial intelligence by the public sector. Technical report. publications office of the European union. https://doi.org/10.2760/39336

Pasquale F (2015) The black box society: the secret algorithms that control money and information. Harvard University Press

Rahwan I (2018) Society-in-the-loop: programming the algorithmic social contract. Ethics Inf Technol 20(1):5–14

Rapeli L (2022) What is the best proxy for political knowledge in surveys? PLoS ONE 17(8):e0272530

Rothstein BO, Teorell JAN (2008) What is quality of government? A theory of impartial government institutions. Governance 21(2):165–190

Saha Debjani, Candice Schumann, Duncan Mcelfresh, John Dickerson, Michelle Mazurek, Michael Tschantz, 2020, "Measuring non-expert comprehension of machine learning fairness metrics." International Conference on Machine Learning. PMLR, pp. 8377–8387

Sandefur J, Birdsall N, Fishkin J, Moyo M (2022) Democratic deliberation and the resource curse: a nationwide experiment in Tanzania. World Politics 74(4):564–609

Sanders Lynn M (1997) Against deliberation. Political Theory 25(3):347–376

Simon N, Veri F, Dryzek JS, Bachtiger A (2023) How deliberation happens: enabling deliberative reason. Am Political Sci Rev 118:345–362

Starke C, Baleis J, Keller B, Marcinkowski F (2022) Fairness perceptions of algorithmic decision-making: a systematic review of the empirical literature. Big Data Soc 9(2):20539517221115188

Thurman N, Moeller J, Helberger N, Trilling D (2019) My friends, editors, algorithms, and I: examining audience attitudes to news selection. Digit J 7(4):447–469

Tyler TR (2021) Why people obey the law. Why people obey the law. Princeton University Press

Waldman A, Martin K (2022) Governing algorithmic decisions: the role of decision importance and governance on perceived legitimacy of algorithmic decisions. Big Data Soc 9(1):20539517221100450

Wang, Ruotong, F Maxwell Harper, and Haiyi Zhu. 2020. "Factors influencing perceived fairness in algorithmic decision-making: algorithm outcomes, development procedures, and individual differences." In: proceedings of the 2020 CHI conference on human factors in computing systems, 1–14

Washington AL (2018) How to argue with an algorithm: lessons from the COMPAS-ProPublica debate. Colo Tech LJ 17:131

Wirtz BW, Weyerer JC, Sturm BJ (2020) The dark sides of artificial intelligence: an integrated AI governance framework for public administration. Int J Public Adm 43(9):818–829

Young IM (2000) Inclusion and democracy. Oxford University Press, Oxford

Zhang Baobao. 2021. "Public opinion toward artificial intelligence." Forthcoming Oxford Handbook of Artificial Intelligence Governance

Zhang B, Anderljung M, Kahn L, Dreksler N, Horowitz MC, Dafoe A (2021) Ethics and governance of artificial intelligence: evidence from a survey of machine learning researchers. J Artif Intell Res 71:591–666

Zhang Baobao, and Allan Dafoe 2019 "Artificial intelligence: American attitudes and trends." Available at SSRN 3312874