Towards Automating Deliberation? The Idea of Deliberative Democracy Embedded in Google's Habermas Machine

Nicolás Palomo Hernández
Scuola Normale Superiore
Sede di Firenze - Palazzo Strozzi
50123 Firenze (Italia)
nicolas.palomohernandez@sns.it

Abstract

Artificial Intelligence (AI) has recently emerged as a central issue for deliberative democracy, both as a subject of deliberation and as a way to improve processes of democratic deliberation. Deliberative democracy scholars and practitioners are already working on the potential of AI-based tools to 'improve', 'enhance', 'augment' or make deliberation more efficient. The present article studies the case of the Habermas Machine (HM), an AI-based tool designed by Google DeepMind to generate agreements among deliberative groups of people with opposing political positions. This article explores the choice of norms and values embedded in the design of the HM and the normative implications of such choices for deliberative democracy. Therefore, the research question guiding this article is: *What are the norms and values embedded in the design of the Habermas Machine, and how do they relate to the normative requirements of deliberative democracy?* The article explores this by addressing how the machine conceives 1) the values for a good deliberation, 2) the space for deliberative politics, 3) the role of humans/participants in deliberation, and 4) the goal of deliberation. The HM has a very narrow conception of deliberative democracy embedded in its design, which 1) overemphasises the desirability of agreement in deliberative processes, 2) limits participation to an individual and private space, and 3) segments and compartmentalises the deliberative process and reduces human participation to the generation and evaluation of political opinions. Instead of empowering and supporting human participation, the normative conception of deliberative democracy embedded in the design of the HM narrows the realm of the political, simplifies the deliberative process, and relegates humans to a secondary role. Building on this specific tool, the article aims to promote and contribute to the conversation between political theory and the work of developers of AI-based tools designed for deliberative purposes. The objective is not simply to guide or orient the design of the technological artefacts, but to evaluate their contributions, limitations, and impacts on the basis of the normative standards on which they are grounded.

Introduction

Technology in general and Artificial Intelligence (AI) in particular have recently emerged as central issues for deliberative democracy, both as a subject of deliberation and as a way to improve processes of democratic deliberation (Goñi 2025). Deliberative democracy scholars and practitioners are already working on the potential of AI-based tools to 'improve', 'enhance', 'augment' or make deliberation more efficient (Landemore 2024; McKinney 2024).

However, the algorithm always implies a simplification and has specific normative preferences embedded (Bucher 2018). It is therefore necessary to research how technologists frame and simplify 'democratic problems' through their designs (Goñi 2025). Careful analysis of how these tools work and relate to deliberative democracy models can help shed light on the foundations of deliberative democracy and what role AI should play in it. This article explores this by studying the case of the Habermas Machine (HM). The HM is an AI-based tool designed by Google DeepMind to promote agreement in collective decisions with individuals holding diverse and opposing political positions. The designers claim that the HM can be helpful for deliberative democracy in general and processes such as citizens' assemblies in particular (Bakker et al. 2022; Tessler et al. 2024).

The present article researches what imaginary of deliberative democracy is embedded in the design of the HM and how it interacts with the broader normative requirements of the theory. Therefore, the question that guides this research is:

> *What are the norms and values embedded in the design of the Habermas Machine, and how do they relate to the normative requirements of deliberative democracy?*

The article begins by exploring the development, functioning, and objectives of the HM and situates it within broader debates on the uses of AI for deliberative democracy. It then examines the norms and values embedded in the design of the HM by addressing how the machine conceives 1) the values for a good deliberation, 2) the space for deliberative politics, 3) the role of humans/participants in deliberation, and 4) the goal of deliberation. Building on this specific tool, this article aims to promote and contribute to the conversation between political theory and the work of developers of AI-based tools designed for deliberative purposes. The objective is not simply to guide or orient the design of the technological artefacts, but to evaluate their contributions, limitations, and impacts on the basis of the normative standards on which they are grounded.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The Development, Functioning, and Objectives of the Habermas Machine

The Uses of AI for Deliberative Democracy

Democracy poses problems that can only be solved through technology (Jungherr 2023). Risse (2022) describes how different democratic ideals have been linked to different technological artefacts, such as the *kleroterion* (the preeminent method for selecting public positions by lot in Athenian democracy) or the ballot box (the central instrument of electoral processes in modern representative democracies). While the former was used to guarantee the neutrality of sortition and ensure the correct composition of the assembly, the latter guarantees anonymity and facilitates the aggregation of individual votes in mass democracies. However, the significance of technology goes beyond its ability to embed and channel the political processes for which it is used. Technology influences the societal fields and processes mediated through it (Winner 2017). For Risse (2022), the 'materiality of democracy' reflects and determines what kind of citizenship and of being human is possible in a democracy. As Goñi (2025, p.9) states, '*publics are not just mobilised, but enacted and performed through devices.*' Artefacts constrain citizens' behaviours and frame and delimit the meaning, functioning, and possibilities of the democratic process for which they are used. As such, the artefacts also become the object of social demands and expectations about democracy. Just as in Athenian democracy the *kleroterion* was expected to promote virtue and citizens' orientation to the common good in the exercise of public office, in mass representative democracies the use of the ballot box is expected to help reveal the will of the political community and to select public representatives who behave accordingly. Therefore, the question is not only how different democratic ideals require different technological artefacts, but also how material artefacts shape and determine the functioning of our democracies.

With AI's development and hype, it seems inevitable that proposals will emerge to encourage its use as a basis for new democratic innovations. Risse (2022) states that '*it might only be a question of time until some techno-optimists propose the large-scale involvement of AI in our collective-choice processes [...] as a way of transferring Aristotle's argument from the wisdom of the multitude*' (Risse 2022, p. 60). In fact, within democratic theory, the subfield that is most willingly exploring the possibilities of AI so far has been deliberative democracy. Deliberative democracy is a talk-centric model of political legitimacy and a theory of democracy in which the association and collective decisions of the political community must be grounded through public argument and reasoning among free and equal citizens (Cohen 2005; Chambers 2009). Due to the constraints to realising this ideal in the mass public, innovative practices of democratic deliberation in controlled environments (e.g. deliberative mini-publics, citizens' assemblies, deliberative polling, etc.) have emerged to promote deliberation among smaller groups of participants who are nonetheless representative of the broader political community (Chambers 2009).

Technology in general and Artificial Intelligence (AI) in particular have recently emerged as central issues for deliberative democracy, both as a subject of deliberation and as a way to improve processes of democratic deliberation (Goñi 2025). Deliberative democracy scholars and practitioners are already working on the potential of AI-based tools to 'improve', 'enhance', 'augment' or make deliberation more efficient. McKinney (2024) lists eleven possible applications of AI in citizens' assemblies, from summarisation and translation tasks to facilitation and generation of consensus statements. Landemore (2024) states that AI can help reconcile the ideals of deliberative democracy and mass participation by scaling up deliberative mini-publics. However, as Goñi (2025) explains, for democratic innovations to become closer to mass participation, the normative values of deliberation would also have to scale up. With the aim of reconciling the ideals of deliberation and mass participation, Landemore (2024) proposes to '*let go the ideal of all minds engaged in one common deliberation and instead settle for an approximation*' in the form of AI-enabled Massive Online Deliberation (M.O.D.) or Many Rotating Minipublics (M.R.P.) (Landemore 2024, p. 40).

As Lechterman (2024) argues, the rise of AI implies an urgent reflection on the foundations of democracy. AI is sometimes assumed to have the potential to change the materiality of democracy. According to Risse (2022), technology is political and determines how the democratic ideal is translated and survives in practice. Risse (2022) states that technology is neither an enemy nor a natural ally of democracy, but can be used for both autocratic and democratic purposes. In the same way that authoritarian regimes use AI to reinforce the authoritarian ambitions of their governments, it would be possible to implement AI design choices that contribute to strengthening democracy (Risse 2022). This view of technology departs from technodeterministic positions and assumes that the design of technological artefacts is flexible and depends on good design choices (Pinch 2012; Winner 2017; Cohen and Fung 2021). For Risse (2022), technological advances should thus be widely debated in democratic politics to ensure that artefact designs are geared to advance rather than undermine democracy. The goal would thus be that decisions about the design and use of technological artefacts reflect the preferences and needs of the groups affected by them, regardless of whether or not they are direct users. This relates to the all-affected principle to guarantee inclusion in deliberation (Mansbridge 2015). In the case of technologies used to facilitate democratic processes, they are not only socially shaped by the preferences of the groups involved in their design but also normatively shaped by the designers' preferences and broader social expectations and democratic claims. Thus, the design and use of technological artefacts used for political processes should be problematised as a political issue to be decided democratically.

Following Jungherr (2023), this article dismisses maximalist assumptions or claims about a fictional overpowered Artificial General Intelligence (AGI). Jungherr (2023) states that social scientists must reflect on the inner workings and domain-specific effects of real and targeted applications of narrow-specific AI, particularly how it is used in politics and by governments. Goñi (2025) argues that a deeper engage-

ment with the material aspects of the democratic process could enrich deliberative democracy. The HM and other AI artefacts for deliberative democracy that are currently being developed provide an opportunity to study narrow-specific applications of AI for political and democratic purposes. Careful analysis of how these tools are designed and relate to deliberative democracy models can help shed light on the foundations of deliberative democracy and what role AI should play in it.

The Habermas Machine as a Technological Artefact

The Habermas Machine (HM) is an AI-based tool designed to substitute human mediation and promote agreement in collective decisions among individuals with diverse preferences. The HM has been developed by Google DeepMind, a research lab specialising in applying AI neural networks, which also develops the popular public-use chatbot Gemini. DeepMind was acquired by Alphabet (Google's parent company) in 2014 due to the tech giant's expansion into AI development. The HM is a Large Language Model (LLM) system designed to facilitate collective decision-making by mediating between humans in deliberative groups. The HM promises to be able to find common ground between individuals with conflicting political positions. The designers of the HM state that the tool can be used as a mediator in deliberative groups in general and, more specifically, in democratic innovations such as citizens' assemblies. The tool has been named after Jürgen Habermas, one of the main contributors to the development of deliberative democracy theory. The designers claim to have made this decision because, they say, Habermas '*proposed that when rational people deliberate under idealised conditions, agreement will emerge in the public sphere*' (Tessler et al., 2024, p.1).

Language Modelling (LM) is an approach used in the field of Natural Language Processing (NLP), a subfield in computer science concerned with providing computers with the ability to process data encoded in natural language (Naveed et al. 2023). Large Language Models (LLMs) are Pre-trained Language Models (PLMs) with significantly larger scales and parameter sizes. PLMs are LMs typically trained via self-supervision on extensive datasets (Naveed et al. 2023). PLMs were developed after the LM's dismissal of Statistical Language Models (SLMs) and the shift of the discipline to subsymbolic AI towards deep learning and neural networks, transforming into Neural Language Models (NLMs) (Naveed et al. 2023). LMs based on deep learning perform better in generalisation and summarisation. LLMs thus constitute a significant advancement from the symbolic conventional statistical LMs, which were trained for task-specific models in supervised settings (Naveed et al. 2023). In short, an LLM could be defined as an AI model designed to process, manipulate, and generate natural language through the recognition and reproduction of patterns in large volumes of text with which they are trained.

In 2022, a group of DeepMind researchers published a paper summarising their progress in creating LLMs to build agreements between humans with diverse preferences (Bakker et al. 2022). Although the term 'preferences' can be understood in a general way, not only concerning political preferences, the proposed questions for the HM-mediated processes were related to major moral or political debates (e.g. '*Should we tax unhealthy foods and sugary drinks?*', or '*Should we re-nationalise the railways?*') (Bakker et al. 2022).

Later, the designers published a paper presenting the tool publicly as the HM and summarising the results obtained in various experiments to test it (Tessler et al. 2024). The political orientation of the HM was explicitly reflected in the paper. The designers openly admitted that the questions for deliberation were related to 'potentially divisive' issues in UK politics (e.g. '*Should we lower the voting age to 16?*' or '*Should the National Health Service be privatised?*') (Tessler et al. 2024). These candidate questions were generated by an LLM from a small set of seed exemplars (Bakker et al. 2022; Tessler et al. 2024). In other words, from a pre-established sample, the AI system itself is in charge of defining the political agenda. After generating the candidate questions, the designers review them and eliminate those '*likely to elicit extremist views or discriminatory language from the users*' (Bakker et al. 2022, p.3) or susceptible to '*provoking offensive commentary*' (Tessler et al. 2024, p.2).

The designers state that HM-mediated deliberation leads to the '*emergence of shared perspectives on potentially polarising social and political issues*' or that the HM could potentially be used for '*conflict resolution, political discussions and citizens' assemblies*' (Tessler et al. 2024, p. 1). While the designers claimed that preliminary models of the tool were intended to produce consensus, DeepMind presented the HM as capable of '*finding common ground*' or '*generating agreement*' (Bakker et al. 2022; Tessler et al. 2024). For practical reasons, I will primarily focus on the most up-to-date model of the system, the HM, as described in Tessler et al. (2024). The previous paper published by the same designers focuses on the development process of the AI system, delving into the process of fine-tuning the pre-trained LLM to improve its performance and adapt it to its final task (Bakker et al. 2022).

The HM has a complex design and functioning. HM substitutes human mediation in collective decision-making and/or processes of democratic deliberation. The HM consists of two complementary LLMs: 1) A generative model that suggests statements that reflect the diversity of positions of the members of the deliberative group, and 2) A personalised reward model that scores statements based on how it believes different members of the deliberative group will agree with it (Tessler et al. 2024). First, participants write (individually and privately) a short statement summarising their position on the initial question. The statements are incorporated into the HM, which, combining its two complementary LLMs, generates a series of group statements according to the predicted preferences of the participants based on the previous individual statements. The group statements are then evaluated by the participants (individually and privately) through a ranked choice election according to the quality of the argument and the level of agreement it reflects with their position. The HM aggregates the results and selects a winning statement. The winning statement is returned

to the participants, who must again write (individually and privately) a short paragraph summarising their position on the statement. The individual statements are fed back into the HM, which generates a series of revised group statements. Again, the participants evaluate the revised group statements through a ranked choice election. The HM aggregates the results and chooses the winning statement, concluding the process of deliberation.

Figure 1 summarises how the HM works; Figure 2 summarises the process by which the HM generates consensus statements.

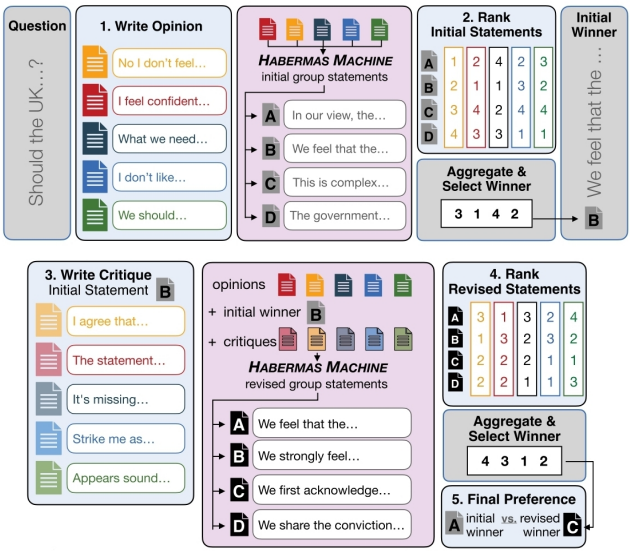

Figure 1: HM's deliberation procedure (Tessler et al. 2024)

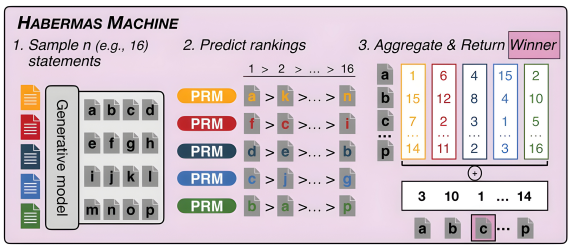

Figure 2: HM's consensus statements (Tessler et al. 2024)

To evaluate the tool, the designers conducted a series of experiments with a sample of 5734 participants from the UK (Tessler et al. 2024). In addition, they replicated their findings using the HM in a virtual citizens' assembly whose participants constituted a representative sample of the UK population (Tessler et al. 2024). To select the representative sample, the designers teamed up with Sortition Foundation, a UK-based company that opposes electoral democracy and advocates the introduction of sortition and deliberative mini-publics for collective decision-making.

The Habermas Machine as a Political Artefact

The HM is an AI-based tool that mediates deliberative groups involving individuals with diverse and opposing political positions. Although the designers argue that there is still much research to be done on the possible risks of such a system before it is implemented (Bakker et al. 2022), the tool has a clear political and public orientation. The designers do not conceal that the tool is intended for this purpose. They highlight its ambitious potential, which could contribute to '*finding agreement and promoting collective action in an increasingly divided world*' (Tessler et al. 2024, p.1).

But why would Google care about deliberative democracy? The fact that Google, one of the world's largest technology companies and the hallmark of 'surveillance capitalism' (Risse 2022), is in charge of designing such an ambitious tool to facilitate deliberative democracy raises clear concerns. Habermas (2022) himself highlights the concentration of market power of large internet corporations as a cause of the disruption of the public spheres of contemporary democracies. Habermas (2022, p. 160) warns of a '*libertarian grimace of world-dominating digital corporations*.'

The HM can be seen as just an output of research labs like DeepMind experimenting with the possibilities of AI and its potential uses for democratic purposes. However, from a broader perspective, the HM would also contribute to shaping citizens' preferences about AI and Google's role in AI development. If the perception of risk associated with AI decreases, so do demands and expectations for AI regulation, which is in line with the interests of the tech giant. Publishing articles on the design, functioning, and objectives of the HM would serve as a marketing communication exercise about a potentially beneficial use of AI for society, with the aim of making the technology more appealing to the general public.

In this sense, the socially beneficial use of AI would be its supposed contribution to promoting agreements and reducing political conflicts through more 'impartial', 'neutral', and 'unbiased' mediation. Building on the experiments to test the tool, the designers even claim that the HM is able to '*overcome the potential for mediator bias*' in collective deliberation by producing less biased statements than human mediators (Tessler et al. 2024, p. 7). However, as Zajko (2022) explains, there is no agreement on the definition of bias. While the conventional view defines bias as '*inaccurate modelling or prediction*', an alternative and broader perspective defines it as '*whatever might be undesirable in society*' (Zajko 2022). As Zajko (2022, p.4) explains, '*unequal ground truth might be considered unbiased in the narrow sense of bias as accuracy, but it may be biased according to this second view of bias as an undesirable tendency.*'

The HM designers claim to have considered 'social benefit' when designing the tool (Bakker et al. 2022). However, there is also no agreement on the desirable use of AI: the definitions of 'social desirability', 'social benefit' or 'public interest' are strictly political. As Bucher (2018) explains, the algorithm always implies a simplification, so it is unable to capture the whole complexity of democratic processes. The algorithm's actions imply political consequences or con-

tribute to producing a particular social order. Bucher (2018) states that algorithms can be understood as '*inscriptions of certain ideologies or particular ways of world-making.*'

For Tasioulas (2022), AI culture is technocratic, as decisions about the values encoded in its designs are made by corporate elites and are outside democratic control. In a democracy, however, it is the people who must decide the future of society and the technological artefacts that materialise it (Coeckelbergh 2023). As Coeckelberg (2023) argues, the responsible development of AI cannot depend on the good faith of the designers but requires policies that push tech companies to create artefacts that are beneficial to democracy. It would be irresponsible, both as citizens and as social scientists, to just assume that the designers are acting in good faith and that the interests of the multinational have no impact on the design of the HM.

It is therefore necessary to research how technologists frame and simplify 'democratic problems' through their designs (Goñi 2025). According to Zajko (2022), critically situating the discourses and imaginary futures that guide the development of artefacts is one of the main tasks that sociologists can undertake to help define the role that AI should play in democratic societies. The HM has a certain normative idea of deliberative democracy embedded in its design. The key issue for political theory is not to study whether the HM is impartial, neutral, or unbiased, but rather what conception of what is socially desirable for deliberative democracy is embedded in its design. In other words, what is the conception of the 'social benefit' or 'public interest' in deliberation that HM designers assume to be neutral but is loaded with political values and preferences. The following section will explore this.

The Norms Embedded in the Design of the Habermas Machine

The present section looks directly at the norms and values embedded in the design of the Habermas Machine (HM) and their interaction with the normative requirements of deliberative democracy. Although the latter would have required a prior presentation of a series of normative standards of deliberative democracy, this has not been possible due to space constraints. Thus, the analysis of the design of the HM will be directly intertwined with references to various normative discussions on deliberative democracy. Many of the criticisms contained in this section extend to other 'orchestrated' practices of democratic deliberation (see Chambers 2009; Hammond 2025). However, through the use of AI, the HM reproduces or exacerbates many of these limitations. This will be explored by addressing how the machine conceives 1) the values for a good deliberation, 2) the space for deliberative politics, 3) the role of humans/participants in deliberation, and 4) the goal of deliberation.

The Values for a Good Deliberation

Tessler et al. (2024, p.7) relate what a good deliberative process should be to three values: 1) Efficiency, 2) Fairness and 3) Scalability.

1. Efficiency: Designers assess the efficiency of the process in terms of time and costs. They argue that, compared to face-to-face democratic deliberation processes, those mediated by the HM have lower costs in terms of time and money. The aim is to build a tool that is easy to implement, easily exportable, and capable of quickly generating outputs (i.e. agreements or decisions). McKinney (2024, p.8) points out that LLMs '*could support participants to find common ground and resolve conflicts more quickly.*' However, to reduce economic and time costs so exponentially, the HM leaves out important phases of the deliberative process and reduces human participation to generating and evaluating political opinions. The goal of making decision-making faster or less costly is obviously not a negative thing *per se*, but it is important to remain wary that it does not conflict with other essential goals of deliberation or limit either the capacity or the depth of the deliberative process. The adequacy of using AI in deliberation should be judged by its ability to hinder or contribute to the democratic legitimacy of the public sphere or to the discursive quality of opinion-formation and will-formation processes (Habermas 2022).
2. Fairness: The definition of 'fairness' for the HM designers follows a utilitarian logic. Bakker et al. (2022) make it explicit when choosing a Social Welfare Function (SWF) that reproduces the utilitarian logic for their reward model for preference prediction. The reward model predicts each participant's level of agreement with each consensus statement generated by the machine and assigns it a value. The machine then aggregates the values of each consensus statement and chooses the one with the highest value, i.e. the one that maximises group welfare or mean expected agreement across the group. This is consistent with Tasioulas' (2022) criticism that data scientists take preference-based utilitarianism as their ethical standpoint. However, if well-being maximisation is assumed to be the fundamental moral injunction and well-being consists in fulfilling human preferences, ethics would be reduced to a prediction and optimisation exercise (Tasioulas 2022). Designers argue that the results of the reward model would be similar if their SWF followed a Rawlsian logic consisting of maximising the expected agreement for the most dissenting group member (Bakker et al. 2022), and even that their model can accommodate the minority opinion and respect the majority opinion at the same time (Tessler et al. 2024). In any case, they claim that the reward model simulates '*a ranked-choice election in which each participant's vote has equal weight*' (Tessler et al. 2024, p.7). However, deliberative democracy does not rely on the maximisation of preferences of the group as a whole or of any of the individuals in the group in particular. Deliberative democracy must ensure a space for reflection on what our preferences should be, instead of '*taking those preferences as ultimate determinants of value*' (Tasioulas 2022, p. 235). The HM conceives preference-formation as external to the deliberative process, and therefore, the HM-mediated process only consists of maximising the preferences of the deliberative group (not on improving the discursive

quality of the arguments or contributing to the participants' formation of political preferences). The HM might encourage participants to reflect on their own preferences individually. However, this reflection is not based on a public and collective process of argumentation, justification, or reason-giving (Chambers 2009), but rather on an algorithmic calculation to generate statements that maximise the support of the group of participants.

3. Scalability: The HM has a greater purpose than facilitating the deliberation of small groups of people. The short-term goal is to bring thousands of people into the process (Tessler et al. 2024). The HM conceives deliberative democracy as a massive but discrete collective dialogue of individual citizens. The scalability intended by the designers would only consist of incorporating a larger number of individual participants through cost reduction and preference aggregation. However, scalability is not only about the inclusion of more participants in the actual deliberative process, but also about the ability of the process to produce systemic effects (Chambers 2009; Mansbridge et al. 2012). The mediation of intermediary collective actors (which are completely side-lined by the HM) is indispensable for the connection between the micro-public and the broader macro-public and for securing the epistemic quality and democratic legitimacy of the public sphere (Olsen and Trenz 2016; Chambers 2023). However, no collective intermediary actors or connection to the broader macro-public is included in any phase of the HM-mediated process. Participation in the process is limited to an individual and private space.

The designers assume that the use and definition of these concepts is neutral and does not imply the adoption of a particular political stance. However, the role given to efficiency, fairness, and scalability to measure the success of the tool reflects specific normative preferences on the part of the designers about what deliberation should aim for.

The Space for Deliberative Politics

I refer to the space for deliberative politics as the margin between the choice of the topics for deliberation and the limits of what is debatable. This would be, therefore, the space of the political, of what is susceptible to being debated.

The political agenda or topics for deliberation in the experimental processes mediated by the HM were AI-generated. From a small set of seed exemplars written by the designers, an LLM generates 'potentially divisive candidate questions', which will guide the HM-mediated processes (Bakker et al. 2022; Tessler et al. 2024). Thus, agenda-setting is left exclusively for AI to produce it building on a sample written by the designers. This is in direct conflict with Dahl's democracy criterion regarding the control of the agenda (i.e., '*members* [of an association] *must have the exclusive opportunity to decide how and, if they choose, what matters are to be placed on the agenda*') (Dahl 1998, p. 37). Innerarity (2023) argues that AI is unable to determine the political problem or the issue at stake on its own, the so-called 'framing problem'. Similarly, Tasioulas (2019) states that AI cannot deliberate on what its ultimate goals should be. AI sets the issues for discussion according to a) a frame previously incorporated by the designers or b) the bias contained in the sample with which it has been trained. In the case of the HM, the AI generates issues based on what designers consider to be 'potentially divisive'. The list of seed exemplars is extensive, ranging from '*Are horoscopes good for people to read?*' to '*Would it be a good idea to further increase taxes for corporations?*' Everything that is introduced in this list shapes the questions that are later generated: from adding the issue of horoscopes (and not, for example, questions related to Big Tech regulation) to the use of words like 'further' to illustrate the increase in taxes (which can imply that taxes have already been raised significantly beforehand). Thus, in generating questions for deliberation, the AI reproduces a 'bias from the past' (Innerarity 2024), extrapolating what is considered political based on what has been considered political in the past. However, AI is unable to generate moments of political rupture or change, underemphasising the intentionality, discontinuity, interruption, or transformation, which are constitutive of democratic politics (Innerarity 2024). By linking what is political to what is 'potentially divisive' as the object of future consensus, AI is unable to identify already established consensuses as objects of possible political deconstruction and debate. The political agenda in a democracy should be set according to what the political community decides, not according to what has traditionally been considered 'potentially divisive'. There may be 'potentially divisive' topics that a political community chooses not to deliberate on; there may be issues that a political community chooses to deliberate on that were not initially considered 'potentially divisive'.

While an AI generates the questions for deliberation, the designers determine the limits of the content of those questions. After the AI generates the questions, the designers filter them and eliminate those that they consider '*likely to elicit extremist views or discriminatory language from the users*' (Bakker et al. 2022, p.3) or susceptible to '*provoking offensive commentary*' (Tessler et al. 2024, p.2). The published material does not specify what the designers mean by 'extremist', 'discriminatory' or 'offensive' and what process they follow to decide which candidate questions should be eliminated. The process appears to be a manual, case-by-case elimination based on the designers' preferences. In addition to concerns about the possible manipulation of the deliberative process, the curation of the issues for discussion by designers jeopardises one of the main requirements of the democratic public sphere, namely, reflexivity. The condition of reflexivity in deliberative democracy entails that nothing should be immune to political contestation, not even the nature of the procedures (Markell 1997). As Innerarity (2023) explains, democracy is a way of organising collective life that politicises everything, where there is nothing impertinent and nothing that cannot be debated. However, by manually filtering the questions for discussion in the HM, the designers define the boundaries of what can and cannot be debated, limiting and narrowing the realm of the political.

Thus, the space for deliberative politics in HM-mediated processes is very narrow and falls between a) the questions generated by the AI (based on a sample of 'potentially di-

visive' questions written by the designers) and b) what the designers consider as likely to be 'extremist', 'discriminatory', or 'offensive'. The political agenda in the HM is co-produced between AI and a very small group of people, the designers of the tool themselves. In any case, the generation of the questions is not exactly part of the HM: the tool could still be used to facilitate deliberative groups on previously decided topics or even decided by the participants themselves.

The Role of Humans/Participants in Deliberation

Risse (2022, p. 72) states that the '*materiality of democracy reflects and determines what kind of democratic citizenship and what ways of being human are available in a democracy.*' The HM's design leaves a very specific role to humans/participants/citizens: generating written individualised opinions and evaluating the machine-generated statements.

The construction and design of the HM is consistent with the labour-based theory of automation put forward by Pasquinelli (2023). Pasquinelli (2023) argues that the design of technological artefacts is shaped by the widely accepted economic principles of the time. According to this view, AI does not imitate biological intelligence, but the structure of labour and social relations (Pasquinelli 2023). For Pasquinelli (2023), the design of technological artefacts mimics the division of labour: different tasks of the same process are segmented to calculate and reduce the costs of each particular task and increase the extraction of surplus value. The labour-based theory of automation plays a central role in the design of new technological artefacts, not only in the pursuit of cost reduction but also in the imposition of labour and social hierarchies (Pasquinelli 2023).

In automating deliberation, the HM segments and compartmentalises the deliberative process. As in the automation of other complex processes, instead of freeing humans from repetitive tasks so they can focus on more creative or intellectually demanding ones, AI limits human intervention to overly simple tasks. In the case of the HM, the role of the participants is limited to intervening on only four occasions during the process: two of them to 'produce' written opinions, and the other two to evaluate the machine-generated statements. Moreover, the participants only give their opinions and evaluate the statements individually and privately. This individual and private participation prevents any argumentative interaction between participants for them to defend or contrast their opinions with their peers, ruling out any deliberative exchange from the outset. The subsequent process is entirely detached from the participants and is left for the machine to handle by aggregating preferences and generating consensus statements. As Risse (2022) points out, the propensity to delegate tasks to AI would be detrimental to human capabilities, especially when facilitating collective decision-making. The HM leaves aside the argumentative or reason-giving skills of the participants or, at least, makes them unnecessary for the processes mediated by the tool.

The realm of the political is significantly reduced by the space given to deliberative politics and by the secondary role to which participants are relegated. Rather than a materialisation of deliberation, the HM becomes a machinery for producing political agreements or decisions (facilitated by the tool's objectives of cost reduction and exportability). The work of the participants is put at the service of the tool and not the other way around. Following Coeckelberg's (2023) differentiation, the participant is a 'political patient' in the HM-mediated process rather than a 'political agent'. Instead of using the HM to nurture the deliberative process and facilitate the deliberative exchange between participants, the information produced by the participants is only used to nurture the autonomous functioning of the machine.

In short, instead of contributing to the flourishing or empowerment of humans/citizens/participants, the HM confers them a minimal place in the process and an isolated, repetitive and secondary task. The unlimited generation of 'potentially divisive' questions by the AI and the productive role to which human participants are relegated could escalate the use of the machine for it to become a tool for mass production of political agreements or decisions, but never a tool for inclusiveness and deliberative democratic legitimacy at the systemic level.

The Goal of Deliberation

The main value guiding the design of the HM is the orientation towards agreement. In early publications on the tool, designers closely related its purpose to its ability to produce consensus, which they conceived of as a '*key pillar of the democratic process*' (Bakker et al. 2022, p.2). However, in the most recent publication, designers have moved away from consensus and explain that the HM is an AI-based device created to promote common ground or agreement between individuals with conflicting political positions (Tessler et al. 2024). Tessler et al. (2024, p.1) define agreement as '*a prerequisite to act collectively*'. It is therefore assumed that agreement is not only 1) possible and 2) desirable for democracy, but also 3) achievable through the mediation of the machine and necessary for its correct functioning. This starting point already implies a normative position on deliberative democracy.

The prominence of orientation to agreement in the design of the machine is reflected in the choice of naming it 'Habermas Machine'. The designers state that they chose to name it after Jürgen Habermas since he '*proposed that when rational people deliberate under idealised conditions, agreement will emerge in the public sphere*' (Tessler et al. 2024, p.1). This, however, is a misunderstanding of the role that agreement plays in Habermas' theory (Habermas 2022). Table 1 summarises the differences between the HM designers' interpretation of Habermas' contributions and the remarks made by Habermas (2022) in his latest article.

Much literature has reproduced a false dichotomy between a consensus-based Habermasian and an agonistic Arendtian view of the public sphere (Markell 1997). However, Markell (1997) argues that Habermas' and Arendt's (or even Mouffe's) visions are not opposed, but complementary. According to Habermas (2022), the form of discourse that governs the public sphere is communicative action, which requires orientation towards agreement among free and uncoerced citizens. However, Habermas (2022) does not im-

Habermas Machine	Habermas (2022)
Algorithmic prediction of participants' preferences	Argumentative exchange of competing positions
Focus on outcome → Agreement as a necessary outcome of deliberation	Focus on procedure → Non-agreement as a possible outcome of deliberation
Agreement as desirable (*telos*)	Orientation to agreement (consensus as 'unattainable')
Political decision as an outcome of a completed deliberative process	Political decision has a presumption of rationality (*Fallibility*)
Discussion topics selected by AI and designers; HM procedures remain uncontested	No topic is immune to political contestation, not even the nature of the procedures (*Reflexivity*)

Table 1: The Habermas Machine against Habermas (Own elaboration, 2025)

ply with this that the communicative action must be ruled by the *telos* or end of agreement, but by the openness of those involved to reach it. Habermas (2022, p. 152) even defines consensus as 'unattainable'. Agreement is understood by Habermas (2022) not as a possible and desirable outcome of deliberation, but simply as a precondition for the kind of discourse that should guide deliberation: there is no alternative discursive category oriented towards 'non-agreement' or disagreement (Markell 1997). If there were no such openness to reach agreements on the part of the participants, but only the will to impose a specific position, the action would not be understood as communicative but as strategic (Habermas 2022). Habermas does not imply that the condition of consensus or agreement is normatively desirable, but just that deliberative procedures are preferable to coercion or violence (Markell 1997).

In contrast, the HM bases its legitimacy exclusively on its output, on the results of the deliberative process it mediates, i.e. on its capacity to produce agreements. The functioning of the HM requires the deliberative process to have an established end. The design of the machine excludes the possibility of the deliberative process ending without an explicit agreement, or with the realisation that the parties involved have irreconcilable positions. By grounding its legitimacy entirely on the output of deliberation, the machine's proper functioning requires the production of a final agreement. If the machine does not produce a final agreement, it means that it has failed or that it is useless. This causes the HM to conceive agreement not only as possible or desirable, but also as necessary in the processes it mediates.

While Tessler et al. (2024, p.7) assume that '*even after deliberation, a plurality of views may persist*', the HM conceives its final agreement or decision as the outcome of a completed deliberative process that is not to be subject to further debate. The final agreement is identified as the result of collective intelligence, as something that 'emerges', i.e. something rational and objective, instead of something to be decided politically. In this way, the HM jeopardises another of the main requirements of the democratic public sphere, namely, fallibility (which requires outcomes to always be open to political contestation). The HM overturns the open-endedness, uncertainty and unpredictability of the future that is constitutive of human freedom in democratic societies (Innerarity 2024). No matter how much data is available or for how long deliberation has taken place, the political decision will always be unclear, refutable and changeable (Innerarity 2023). The HM reproduces a narrow but dominant framework of deliberative democracy, which overlooks democracy as a game of shifting and divergent loyalties and identities (Olsen and Trenz 2016). Not everything is resolvable, and not every interest can be accommodated in a political decision (Innerarity 2024). The HM leaves aside agonistic or conflict-prone conceptions of the democratic process, which understand conflict as unavoidable and constitutive of the political (see Mouffe 1999; Arendt 2006). However, the latter is also very much present in Habermas' writings, despite the designers using Habermasian theory to justify the whole design and functioning of the HM. For Habermas (2022), democracy should not *per se* generate agreement or consensus but should provide adequate institutions to test the validity of political statements. Habermas (2022, p. 152) states that deliberative politics enable us '*to improve our beliefs through political disputes to get closer to correct solutions to problems.*'

However, Habermas' deliberative proceduralism assumes that the outcomes of deliberation are fallible and are always open to contestation (Markell 1997). Habermas (2022) does not assume the possibility or desirability of a rational consensus as an outcome of deliberation, but only states that the outcome of the deliberative process has a 'presumption' of rationality. Any agreement, even the rules of the game, must be open to political contestation (Markell 1997). In Habermas, 'orientation towards agreement' is a strong normative claim regarding the processes that make agreement possible, but a weak normative claim regarding the condition of agreement itself (Markell 1997). Thus, Habermas's 'orientation towards agreement' not only admits, but requires contestatory political action (Markell 1997).

In order to produce or arrive at an agreement, the HM even leaves aside human-to-human argumentative interaction. This is paradoxical, as designers justify their emphasis on agreement because they assume that it is '*a prerequisite to act collectively*' (Tessler et al. 2024, p.1). However, human participation through the HM is confined to an individual and private space, preventing any collective action or argumentative deliberative interaction between participants. Participation in the HM consists of writing individual and private opinions and ranking the degree of agreement with the statements generated by the HM. Not only does this prevent the possibility of agreement-oriented com-

municative action, but it also prevents any argumentation and reason-based communication between participants. The HM does not allow participants to take any contestatory political action or even to contrast their opinions and arguments in a deliberative exchange. The designers state that in HM-mediated processes, participants are able to change their opinion to one shared more widely by the whole group (Tessler et al. 2024). However, such a change of position cannot be explained by an argumentative communicative exchange between peers, but by the machine's imposition of a potential agreement based on an algorithmic prediction of what the participants will consider more or less 'acceptable'.

Discussion and Conclusions

The present article has studied the case of the Habermas Machine (HM), an AI-based tool designed to substitute human mediation and generate agreements in deliberative groups composed of participants with opposing political positions. This article has explored the choice of norms and values embedded in the design of the HM and the normative implications of such choices for deliberative democracy.

The case of the HM illustrates how AI artefacts for deliberative purposes have embedded normative preferences in their design and contribute to reproducing specific imaginaries of democracy. However, technological artefacts are flexible and socially malleable and can be made available to democracy through good design choices (Pinch 2012; Winner 2017; Cohen and Fung 2021). The fact that the development of the HM was not open to wider audiences may explain why the tool has such a narrow conception of deliberative democracy embedded in its design, which 1) overemphasises the desirability of agreement in deliberative processes, 2) limits participation to an individual and private space, and 3) segments and compartmentalises the deliberative process and reduces human participation to the generation and evaluation of political opinions. Instead of empowering and supporting human participation, the HM narrows the realm of the political, simplifies the deliberative process and relegates humans to a secondary role. Moreover, HM-enabled processes are open to manipulation and fail to meet two basic requirements of the democratic public sphere: reflexivity (i.e. nothing should be immune to political contestation, not even the nature of the procedures) and fallibility (i.e. outcomes must always be open to further political contestation) (Markell 1997).

As Goñi (2025) explains, the normative values of deliberation would have to scale up for democratic innovations to become closer to mass participation. In this respect, Landemore (2024) has proposed that, to reconcile the ideals of deliberation and mass participation through AI, we must '*let go the ideal of all minds engaged in one common deliberation and instead settle for an approximation*' (Landemore 2024, p. 40). However, using the HM would require lowering our standards for deliberation or deliberative democracy to a minimum. As McKinney (2024, p.4) explains, when integrating AI in deliberation, the technology '*should be used to empower and support human participation, rather than replace it.*' A process where the only role played by human participants is limited to giving individualised and private opinions without any argumentative exchange between them is not a deliberative process. The HM illustrates the risk that some devised technological solutions for the 'problems' of deliberative democracy may actually undermine its core principles (Goñi 2025).

The HM simplifies deliberation to a process aimed solely and exclusively at finding common ground and producing agreements. If the HM does not produce the output of an agreement, it means that it has failed or that it is useless. This makes agreement not only possible and desirable but also necessary in the processes where the HM is used. In contrast, the validity of democratic deliberation is reflected in its contribution to the democratic legitimacy of a political system, regardless of whether it culminates in an agreement. Deliberation does not make agreement a necessary outcome, and therefore does not fail if agreement is not reached. Even if an agreement is not reached, deliberation can reinforce the popular authorisation and democratic legitimacy of the public sphere and the discursive quality of the processes of opinion-formation and will-formation (Habermas 2022). The use and design of technological artefacts for democratic purposes should not be oriented towards the simplification of processes or the generation or production of agreements, but rather towards reinforcing the popular authorisation and democratic legitimacy of the public sphere or helping citizens navigate its plurality and complexity.

AI developers currently demand humanistic and sociological normative approaches to their artefacts. Computer engineers read and reference political theory to guide and justify their designs. Our obligation as social scientists is to join the conversation. Thus, this article has sought to promote and contribute to the conversation between political theory and the work of developers of AI-based tools designed for deliberative purposes. The objective is not simply to guide or orient the design of the technological artefacts, but to evaluate their contributions, limitations, and impacts on the basis of the normative standards on which they are grounded. Exploring if there are AI designs that could be beneficial for deliberative democracy requires us to be open to the possibility that the answer is a negative one. This entails ruling out any temptation of 'technological solutionism' for democratic citizen participation (Goñi 2025). As Zajko (2022, p.7) explains: '*technologists and a broader public need to be regularly reminded that just because we can build or automate something, does not mean that we should, and that rejection of the latest 'innovation' in social control is a perfectly reasonable option.*'

In any case, this article has several limitations. The HM has not been made available to the public and has only been used in experimental tests controlled by the design team itself. Therefore, this article does not analyse the practical functioning of the tool, but only what the Google DeepMind designers have published and disclosed about its development and design. Future research should monitor the development of the HM and supplement the analysis of the values implicit in the secondary information on the tool's design with primary information, such as interviews with designers and empirical analyses of the tool's use in practice.

References

Arendt, H. 2006. *On Revolution*. Penguin Classics.

Bakker, M.; Chadwick, M.; Sheahan, H.; Tessler, M.; Campbell-Gillingham, L.; Balaguer, J.; McAleese, N.; Glaese, A.; Aslanides, J.; Botvinick, M.; et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in neural information processing systems*, 35: 38176–38189.

Bucher, T. 2018. *If... then: Algorithmic power and politics*. Oxford University Press.

Chambers, S. 2009. Rhetoric and the public sphere: Has deliberative democracy abandoned mass democracy? *Political theory*, 37(3): 323–350.

Chambers, S. 2023. Deliberative democracy and the digital public sphere: Asymmetrical fragmentation as a political not a technological problem. *Constellations: An International Journal of Critical & Democratic Theory*, 30(1).

Coeckelbergh, M. 2023. Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*, 3(4): 1341–1350.

Cohen, J. 2005. Deliberation and democratic legitimacy. In Matravers, D.; and Pike, J., eds., *Debates in Contemporary Political Philosophy: An Anthology*, 2352–370. Routledge.

Cohen, J.; and Fung, A. 2021. Democracy and the digital public sphere. In Bernholz, L.; Landemore, H.; and Reich, R., eds., *Digital technology and democratic theory*, 23–61. University of Chicago Press.

Dahl, R. A. 1998. *On democracy*. Yale University Press.

Goñi, J. 2025. Citizen participation and technology: lessons from the fields of deliberative democracy and science and technology studies. *Palgrave Communications*, 12(1): 1–13.

Habermas, J. 2022. Reflections and hypotheses on a further structural transformation of the political public sphere. *Theory, Culture & Society*, 39(4): 145–171.

Hammond, M. 2025. Deliberative democracy without deliberation. *Journal of Deliberative Democracy*, 21(4).

Innerarity, D. 2023. Predicting the past: a philosophical critique of predictive analytics. *IDP. Revista d'Internet, Dret i Política*, 39: 1–12.

Innerarity, D. 2024. The epistemic impossibility of an artificial intelligence take-over of democracy. *AI & society*, 39(4): 1667–1671.

Jungherr, A. 2023. Artificial intelligence and democracy: A conceptual framework. *Social media+ society*, 9(3): 20563051231186353.

Landemore, H. 2024. Can artificial intelligence bring deliberation to the masses. In Chang, R.; and Srinivasan, A., eds., *Conversations in philosophy, law, and politics*, 39–69. Oxford University Press.

Lechterman, T. M. 2024. The perfect politician. In Edmonds, D., ed., *AI morality*, 53–63. Oxford University Press.

Mansbridge, J. 2015. A minimalist definition of deliberation. In Heller, P.; and Vijayendra, R., eds., *Deliberation and development: Rethinking the role of voice and collective action in unequal societies*, 27–50. World Bank Group.

Mansbridge, J.; Bohman, J.; Chambers, S.; Christiano, T.; Fung, A.; Parkinson, J.; Thompson, D. F.; and Warren, M. E. 2012. A systemic approach to deliberative democracy. *Deliberative systems: Deliberative democracy at the large scale*, 1–26.

Markell, P. 1997. Contesting Consensus: Rereading Habermas on the Public Sphere. *Constellations: An International Journal of Critical & Democratic Theory*, 3(3).

McKinney, S. 2024. Integrating artificial intelligence into citizens' assemblies: Benefits, concerns and future pathways. *Journal of Deliberative Democracy*, 20(1).

Mouffe, C. 1999. Deliberative democracy or agonistic pluralism? *Social research*, 745–758.

Naveed, H.; Khan, A. U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; and Mian, A. 2023. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Olsen, E. D.; and Trenz, H.-J. 2016. The micro–macro link in deliberative polling: science or politics? *Critical Review of International Social and Political Philosophy*, 19(6): 662–679.

Pasquinelli, M. 2023. *The Eye of the Master: A Social History of Artificial Intelligence*. Verso Books.

Pinch, T. 2012. The social construction of technology: A review. In Fox, R., ed., *Technological change*, 17–35. Routledge.

Risse, M. 2022. Artificial Intelligence and the Past, Present, and Future of Democracy. In Voeneky, S.; Kellmeyer, P.; Mueller, O.; and Burgard, W., eds., *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives*, 85–103. Cambridge University Press.

Tasioulas, J. 2019. First steps towards an ethics of robots and artificial intelligence. *Journal of Practical Ethics*, 7(1): 49–83.

Tasioulas, J. 2022. Artificial intelligence, humanistic ethics. *Daedalus*, 151(2): 232–243.

Tessler, M. H.; Bakker, M. A.; Jarrett, D.; Sheahan, H.; Chadwick, M. J.; Koster, R.; Evans, G.; Campbell-Gillingham, L.; Collins, T.; Parkes, D. C.; et al. 2024. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719): eadq2852.

Winner, L. 2017. Do artifacts have politics? In Weckert, J., ed., *Computer ethics*, 177–192. Routledge.

Zajko, M. 2022. Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, 16(3): e12962.