DEPOLARIZING AND MODERATING SOCIAL MEDIA WITH AI

Pedro Ramaciotti

ie UNIVERSITY
CENTER FOR THE GOVERNANCE OF CHANGE

AI4 DEMOCRACY

This image was created using the AI tool Adobe Firefly.

DEPOLARIZING AND MODERATING SOCIAL MEDIA **WITH AI**

JULY 2024

DEPOLARIZING AND MODERATING SOCIAL MEDIA WITH AI: TOOLS AND GUIDELINES LEVERAGING REPRESENTATION SPACES

The emergence of a public space hosted on online platforms has seen the appearance of algorithms as mediators that filter, curate, and select the contents we see as users.

As the entanglement between this digital public sphere and offline politics progresses, several works have addressed concerns regarding online segregation and polarization in social media. With a consensus settling on the existence of a global process of democratic decline connected with polarization and with the lack of trust in representation and institutions, fears regarding the role of social media in this process have also expanded. This article discusses a particular technical opportunity presented by the rise of AI systems as mediators. This opportunity leverages a parallelism between two theoretical and seldom connected perspectives:

1) spatial models of politics arising in political sciences, and

2) representation learning spaces ubiquitous in recent and growingly ubiquitous forms of AI.

By laying a bridge between the two models, this article proposes that the emergence of AI also enables the possibility of delimiting and disentangling the part of computation related to politics (and potentially involved in political segregation and polarization), opening a path towards better tools for social platform and AI compliance, regulation, and design tools.

WRITTEN BY
PEDRO RAMACIOTTI

Complex Systems Institute of Paris Ile-de-France (ISC-PIF CNRS), médialab Sciences Po, and LPI Université Paris Cité Paris, France

CONTENTS

1. INTRODUCTION

Artificial Intelligence (AI) refers to a set of computational techniques and systems capable of performing—both simple and complex—tasks historically reserved to human cognition. Depending on the delimiting criteria used, AI systems have existed for decades or centuries. The recent debate surrounding AI, however, is best framed in light of advances in statistical learning in the 2010s and 2020s, but also in view of how they have captured the attention of the public, bringing about with them a renewed debate about the role and the impact of computation in society and democracy.

This article proposes an introductory examination and a structure of this debate in regards to the role of AI in the—digital—public sphere, and its impacts on politics and democracy. From this framework, this article will discuss how recent advances in AI present novel opportunities for understanding and improving their role as mediators. A central claim of this article is that the prevalence of AI systems relying on recent forms of spatial representation learning enables the quantitative assessment and the formalization of the degree to which AI systems rely on and may impact political dynamics such as online segregation and polarization.

THE AI DEBATE.

The pervasiveness of algorithmic computation in society has raised several commentaries ranging from optimism to negative omens.

> The automation of tasks previously performed by individuals, for instance, may free resources such as time, but fundamentally change the value of labor
>
> (MCGAUGHEY 2022).

Computational assistance of tasks performed by humans, may also increase outputs or even render tasks achievable for individuals who previously lacked the sufficient skills. An illustrative example is provided by AI systems designed to assist in software development by proposing code to programmers and that auto-complete lines during writing (Dakhel et al. 2023). AI systems have also made strides in numerous research fields, highlighting the distinction between

scientific knowledge as information on the state of the world and the understanding of its mechanisms (Krenn et al. 2022). The transformative nature of computation reaches arguably into all dimensions of society, often in connection with complex systems and dynamics (*e.g.*, scientific research, the economy, politics, democracy), raising a large number of commentaries regarding the positive and negative consequences of AI. Within this multifaceted debate and the body of academic works addressing it, this article is specifically concerned with the impact of AI systems in the public sphere, and with phenomena such as political polarization and segregation.

DELIMITING AI SYSTEMS.

In computer sciences, the term AI is often meant to refer to neural network architectures (Bishop 1994), while debates regarding policy and regulation adopt a broader scope (*e.g.*, the all-encompassing definition adopted AI Act of the European Union; AI Act 2021). This article will consider first a broad and inclusive definition of AI to detach from the debate on what intelligence is and whether it can be emulated by machines, and to minimize the dependence of its claims on the pace of advancements of the state of the art. Building on this very broad scope, we will then focus on a particular family of computations—based on representation learning—to identify how they provide an opportunity in measuring the impact algorithms in political dynamics and designing new tools for compliance and moderation for online platforms, and more generally AI-driven services.

This article detaches itself from the debate over the general impact of AI in society to focus on the impact of computation on social platforms and politics, laying out the theoretical background needed to assess a technical opportunity here put forward. In doing so, the following sections will develop the necessary theoretical framework to articulate research in computer and political sciences with policy and practitioner communities. Additionally, this article is concerned with a particular domain in which the impact of AI has attracted a wealth of works: the raise of social media to prominent arenas in the public sphere, the role of algorithms in their mediation role through filtering and recommendation, and their broader effects on politics and democracy.

AI SYSTEMS MEDIATING THE PUBLIC SPHERE.

> The emergence of digital platforms as arenas of public discussion has had a profound transformation in the functioning of the public sphere.
>
> (JANSSEN AND KIES 2005)

These platforms lower social and technical barriers to the production and circulation of information: virtually anyone can create contents to compete (at least in principle) next to those produced by established media or political figures. These contents may circulate obeying a logic that disentangles reach with the epistemic power of traditional gatekeepers (Shoemaker and Vos 2009; Bennett and Pfetsch 2018). But because of this new relative abundance of information, and given the limited attentional resources of the public, effective navigation and participation in these arenas mandates algorithmic assistance (*i.e.*, information retrieval, filtering and recommendation), most familiarly embodied in the algorithmic curation of social media *feeds* or *walls*. While every information creation and dissemination process (including those predating the internet) is subjected to a socio-technical context (Barzilai-Nahon 2009), algorithmic mediation is novel in its level of institutional and technical concentration: arguably, a few lines of code controlled by a handful of companies are central in deciding which information is shown to each social media user[1]. Several research works have shown that social media may serve as a news provider (Kwak et al. 2010), thus fulfilling a critical role in the public sphere, a trend observed across an increasing number of countries (Kalogeropoulos et al. 2019).

THE RISKS OF AI MEDIATION IN DIGITAL PLATFORMS.

These changes to the public sphere have met both enthusiastic and pessimistic commentaries. While some scholars have highlighted the potential positive role of the democratization of information and coordination (Shirky 2009), others have sought to moderate such expectations, pointing out at how offline social structures permeate these online spaces (Morozov 2011). Along with the emergence of digital public arenas on the internet, several concerns were raised pointing to the risks of political segregation and polarization they might foster. A prominent family of concerns relates to the level of personalization that these new settings allow through selective

1 As of 2023, 59% of EU individuals were social media users (EUROSTAT 2024).

exposure. As each individual sets up its own informational environment (*e.g.*, though subscription to content producers, or selecting its social network of friends), one identified risk was that of the public becoming "balkanized" or segregated in groups of like-minded people, driven increasingly apart in their political views and thus exacerbating polarization (Sunstein 2001). While some degree of polarization is a natural feature of democracy, these narratives point to the risk of detrimental levels of polarization hindering societal coordination (Vasconcelos et al. 2021) or inciting violence (Feinberg et al. 2022). The prevalence of AI systems mediating these environments, because they can sustain or increase personalization in information consumption, spell additional risks for segregation, popularized among the public as *filter bubbles* (Pariser 2011): groups of individuals informationally segregated by virtue of how they arrange their local digital environments, but *also* because algorithms fail to recommended diverse contents capable of breaking segregation.

These risks and their visibility have driven a large body of scientific works spanning for more than two decades, from the analysis of segregation in online blogging (Adamic and Glance 2005), to mainstream social platforms such as Twitter (Huszár et al. 2022) and Facebook (Guess et al. 2023), and to the impact of generative AI services (Rozado 2023; Liu et al. 2022). The picture emerging from these works is, however, mixed, defying popular narratives. Depending on the setting (social platform, country, period, and population, among other factors), these AI-mediated social platforms have been found to both increase and decrease the diversity of content to which users are exposed (see the work of Lorenz-Spreen et al. 2023 for a systematic review on causal mechanisms). Moreover, the assumption that political segregation (driven either by how individuals configure their local digital environments or by algorithmic recommendations) increases polarization has been also contested. One counter example to the narrative linking polarization to segregation is provided by the work of Bail et al. (2018). In this work, an experiment paying Twitter users to follow cross-cutting political content in the US (*e.g.*, proposing Liberal-leaning content to Conservative-leaning individuals) showed that an increase in political diversity of consumed content can exacerbate polarization instead of moderating it.

Today, there is no clear-cut answer to the question for the role of AI in segregation (in part because of the diversity of online settings), nor a sufficiently general understanding of the link between segregation and polarization.

DESIGNING BETTER ALGORITHMS.

Despite these unknowns, and because of the reach and relevance of the digital public sphere, social platforms and providers of AI services must constantly monitor and improve algorithms with regards to segregation and polarization (among other concerns). A tradition within the recommender systems community in computer sciences has framed this problem as the need to diversify recommendations (Ziegler et al. 2005), specially in a trade-off between diversity and accuracy (Zhou et al. 2010), providing content that the user will engage with, thus sustaining the economic model of platforms. Concretely, this problem has taken two forms in algorithm design: either optimize a utility function integrating both accuracy and diversity, or setting constraints in minimal content diversity that algorithms should propose and then optimizing for accuracy.

> **This article explores how the current state of the art in AI presents new opportunities for addressing the risks of algorithmic mediation in online social platforms. This exploration makes a theoretical connection between two fields seldom articulated in AI research: representation learning in computer sciences and spatial models of political opinions in political sciences.**

Representation learning (also called feature learning) refers to a family of methods in machine learning in which AI systems first perform a feature characterization of input data to then perform targeted tasks (such as prediction, regression, ranking, or classification) on the basis of these features (Bengio et al. 2013). In this family of methods, input data such as online content taken as inputs for the computation of recommendations, are represented in feature spaces. Recommendation tasks (*e.g.*, friend or content recommendations) are then computed on the basis of proximity (*e.g.*, recommending friends to follow whenever they are close in feature space). Most algorithms rely on representation of inputs on abstract *latent* spaces without attributed spatial semantics: *i.e.*, along dimensions of features with no explicit human-intelligible meaning. In such cases, the attribution of semantic meaning to these spatial representations using reference data (a task understood within the field of AI explainability; Burkart and Huber 2021), takes importance, as it improves the evaluation, understanding, and accountability of these AI systems. This ubiquitous family of computations includes, for instance, matrix factorization (Luo et al. 2014) and transformer architectures (Chen et al. 2019); two examples

in which representation space is not readily attributable with meaning by design. Spatial models in political science, on the other hand, are methodological and conceptual tools inherited from political economy approaches (Downs 1957) used in explanations of political behavior. In spatial models, entities considered in a research design—individuals and parties (Jolly et al. 2022), representatives, bills (Poole and Rosenthal 1985), and even news content (Bakshy et al. 2015)—are attributed a position in a space in which dimensions stand as indicators of positions towards political issues. By making a theoretical connection between these two fields, this article proposes that the current state of the art in AI systems allows to recast the problem of algorithm design for managing risks associated with selective exposure in a novel way. Linked to the field of AI explainability, this theoretical connection has the potential of casting this AI security problem by changing the focus from a *normative* (How much diversity should be prescribed in recommendations) to an *independence* problem:

> How to compute recommendations in a way that ignores the political dimensions of a given digital arena?

From a disciplinary perspective, a main objective of this article is to propose an articulation at the interface between computer science, political sciences, and policy and regulation communities, and on which to predicate this new AI security problem.

This article is structured as follows.

- It will first present a discussion on the delimitation of AI with regards to its impact in the digital public sphere, framing it in the broader debate of the impact of AI in democracy.
- Then, a brief examination of the notion of political spaces will be proposed, followed by a presentation of some notions of results in AI explainability, making the theoretical link between both fields.
- Finally, the concepts explored here will be illustrated with a data-driven example, highlighting the challenges ahead in connecting the state of the art in AI with new toolkits and guidelines for algorithmic design.

Building on the theoretical connection developed in this article, on the discussion of recent works in AI explainability, and on the data-driven example presented, a list of guidelines and challenges are presented as a proposition in exploiting these opportunities given the current state of the art.

2. ARTIFICIAL INTELLIGENCE, THE PUBLIC SPHERE AND DEMOCRACY

2.1 WHAT IS AND WHAT ISN'T AI?

While the history of these techniques and systems may date back centuries depending on the criteria used to evaluate which human cognition is task is being emulated, AI is most recently discussed in light of advances in statistical machine learning achieved at the turn of the century, but also in view of how they have captured the attention of the public. These advances and changes in perceptions have brought about with them a renewed debate about the role and the impact of computation in society and in particular in democracy.

When examining the question of the impact of AI in the public sphere and democracy, it is crucial to set the boundaries of the objects to which this term makes reference. In this regard, both the scientific literature and the public debate experience varying degrees of ambiguity, and for good reasons. Several researchers and practitioners reserve the term AI to refer to deep-learning systems or neural network architectures (LeCun et al. 2015)[2], while important regulatory, policy and governmental instances propose an all-encompassing definition[3]. The former form is preferred in discussions aimed at distinguishing system designs and how they perform with respect to the state of the art (*e.g.*, in metrics of accuracy for tasks such as classification or regression), while the latter is preferred when structuring debates regarding risks involved in computation at scale in society and adequate safeguards. Throughout this article we will adopt the latter definition for the sake of completeness, to then focus on the vast but delimited family of computations based on representation learning in order to identify how they provide an opportunity in measuring the impact algorithms in political dynamics. This article also detaches from the debate surrounding the comparison between artificial and human intelligence, to treat AI as a description of a set of computational procedures. The claims put forward in this document do not rely on how AI systems compare to human cognition and capacities, and are thus removed from the debate surrounding the notion of intelligence itself[4].

2 The reader is referred to the work of Norvig and Russel (2002) for an exposition involving a broader scope of AI systems.

3 The definition provided by the AI Act of the European Union (Annex I, AI Act 2021) encompasses machine learning approaches, logic- and knowledge-based approaches (including symbolic approaches), inductive procedures, any use of knowledge bases, inference, mechanic deductive procedures.

4 An example of such a debate is the question of the ill-defined *Artificial General Intelligence* as measurable against human capabilities.

The inclusive definition of AI such as that provided in the AI Act of the European Union (Annex I) is structured along three categories:

1. **Machine Learning.** This type of definition includes supervised, unsupervised, and rein- forcement learning using a wide variety of methods including neural networks and deep learning.
2. **Symbolic approaches.** This second type of definition includes logic- and knowledge-based approaches, including knowledge representation, inductive logic or programming, knowledge bases, inference, deductive and rule-based engines, symbolic reasoning and expert systems.
3. **Statistical approaches.** This third type of definition includes statistical approaches, Bayesian estimation, search and optimization methods, including regression analysis. This third type of definition exists as separate from Machine Learning approaches in part to accommodate in the latter the possibility of approaches that do not rely on any statistical framework of theory.

It is worth noting that the set of computational procedures encompassed by very comprehensive defini- tions include most systems, from spreadsheets and rudimentary arithmetic processors to sophisticated deep-learning systems that rely on massive amounts of data and computation. This inclusive scope has notably sparked debate on the feasibility of operationalizing recent AI regulation, and on the effects that such regulations would have on the economy and innovation (Buocz et al. 2023). Similar to the role it plays in structuring debate on regulation, adopting this initial set of definitions removes the dangers of the substantialism involved in discussing *intelligence*.

The relevance of the notion of AI in the opportunities for moderation outlined in this article lies along three properties, which will further narrow the scope of systems in following sections: accuracy, popularity, and the capacity to produce spatial representation.

ACCURACY.

First, the recent scientific achievements in machine learning and statistical approaches have widened their applications. Proposing a comprehensive overview of the evolution of the accuracy of AI systems in performing different cognitive tasks is not possible without also providing an overview of this landscape of tasks targeted by AI designers. As such, no clear-cut and quantifiable general assertions exist on the progression of the accuracy of AI. The examination of the advances with regards to a few of these tasks, however, illustrates

how improvements on accuracy explain not only the widespread adoption of these systems, but also the renewed interest in them and the increasing importance that they are set to play in the future. Scholars in the sociology of sciences sometimes point to recent advances in computer vision as a triggering use case driving the current boom of AI, or *AI spring* (Bommasani 2023). In this field of machine learning, a pivotal and illustrative moment is provided by the success of neural network approaches in image classification achieved in the early 2010s (Krizhevsky et al. 2012)[5]. More generally across AI tasks, different metrics have seen standards progress from 60% to 90% accuracy, although with large disparities not without caveats regarding the different nature of these tasks and diverse types of accuracy metrics (see the work of Martínez-Plumed et al. 2021 for an overview of these metrics and their evolution).

POPULARITY.

This increase in accuracy across different cognitive tasks is one of the main factors explaining the popularization of AI systems in science and industry: more accurate systems are more reliable, they allow to anticipate the quantifiable quality of the outcome[6]. A second important factor in the popularization is the flexibility with which newer AI systems can integrate in their computations data of different types. In the 2000s and early 2010s, the data provided as input for recommender systems needed to be formatted according the specifications of the algorithm used: *e.g.*, in collaborative filtering (Schafer et al. 2007), as user-user or user-item matrices representing who had interacted with whom or who had read or clicked which content in the past, in order to compute recommendations (Bobadilla et al. 2011). More recent AI systems leverage encoding of heterogeneous forms of data and rely on representation of these different data on common spaces, making *multimodal*[7]: text, images, networks, and even sound and video can be embedded or encoded into a common space to then performs computations used in recommendation, such as ranking by proximity. Combined, these factors have led a trend that has seen several industries and fields rely on the same ubiquitous process, consisting in a phase of embedding input data, followed by a phase of recommendation (and other downstream tasks such as regression, generation, classification).

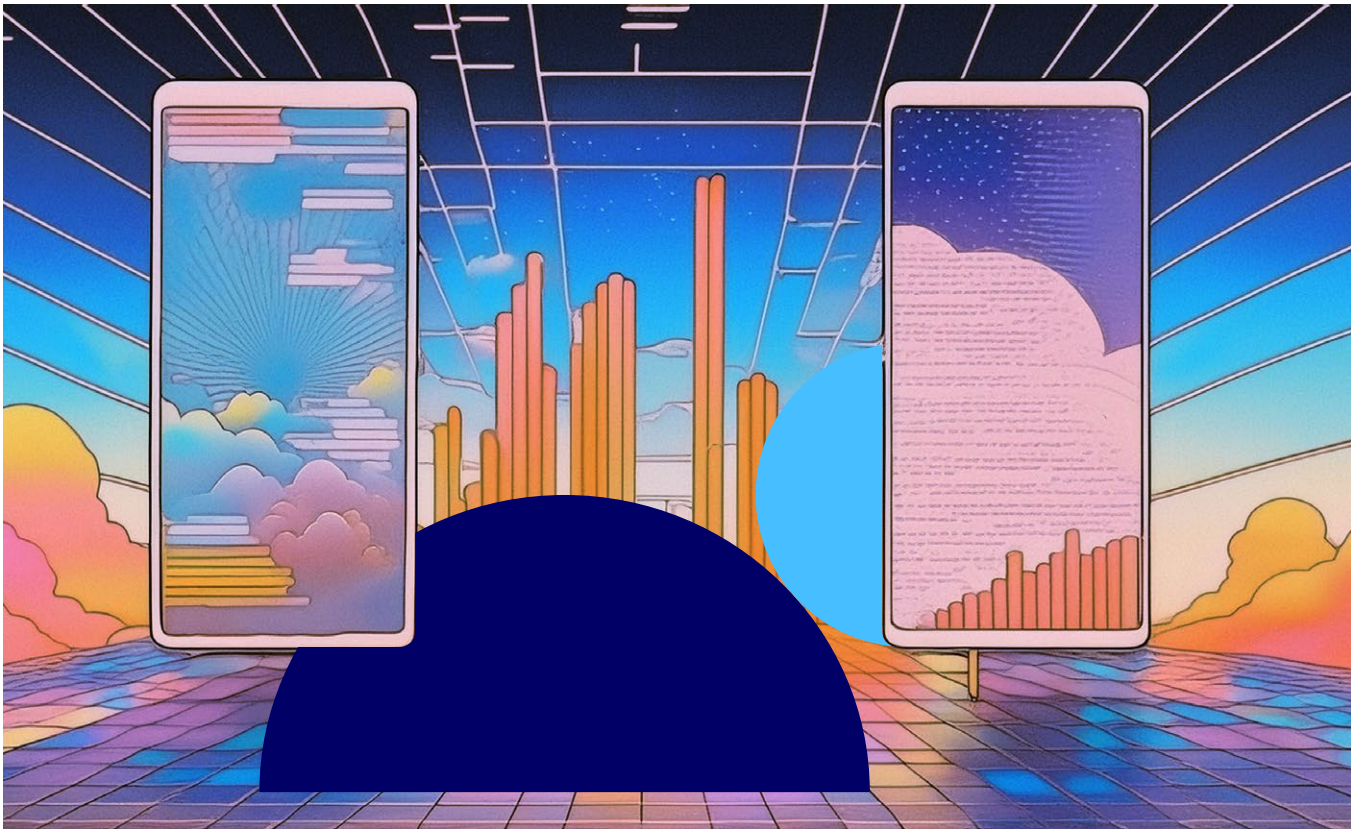
5 At the 2012 European Conference on Computer Vision, Hinton's group presented a neural network approach beating accuracy metrics achieved with traditional, nearly halving the classification error-rate in benchmarks used in competition in the field (Cardon et al. 2018).

6 This quantifiable quality is related to the benchmark metrics on which systems are measured during training and evaluation. This presentation excludes the consideration of the question of the evaluation in generalization of applications outside the data with which the system is trained and evaluated.

7 Traditionally, this is referred to as Multimodal Machine Learning in the AI community (Baltrušaitis et al. 2018).

SPATIAL DATA REPRESENTATION.

The third property of some AI systems that is relevant for opportunities for improving moderation tools for the digital public space is that displayed by the representation learning family of systems. In this family of methods, AI systems first perform a feature characterization of input data in a so-called encoding or embedding phase. Using these representations, tasks such as prediction, regression, ranking, or classification are computed as operations on the spatial positions of the input data on these spaces (Bengio et al. 2013). This family of computations, has been in use since the early 2000s, with algorithms such as Matrix Factorization (Luo et al. 2014), but has found widespread adoption through neural network architectures such as transformers[8] (Vaswani et al. 2017), which have an encoding phase in which they embed input data into multidimensional spaces. This logic of spatial embedding is also widespread in AI mediating social media. Twitter's recommender system, for instance, relies on several core algorithms, such as SimClusters (Satuluri et al. 2020) a general purpose representation algorithm capturing a user's affinities for topics as interpretable vectors of a space on which recommendations are computed. It must be noted that interpreting or explaining the



8 Transformer architectures in neural networks and deep learning systems are central to the acceleration of the interest and adoption of recent generative AI systems, such as ChatGPT, which stands for Generative, Pretrained Transformer.

spatial meaning of these spaces is in general a challenging task. Concretely, this property delimits a widely used family of AI systems in which heterogeneous input data is first represented in abstract multidimensional spaces, on which tasks (including recommendation) are performed as geometrical operations: *e.g.*, ranking according to distances, angular similarities, clustering and classifying by regions of space, or predicting values akin to multivariate regression.

In summary, the scope of AI systems considered in this article starts with the most widely definable set of computational procedures so as to avoid a reliance on the debate of whether AI is comparable to human cognition.

Starting from that scope, this article further delimits the set of AI systems in consideration to those that are capable of taking heterogeneous inputs producing separate computations in two phases:

1. encoding/embedding of input data in abstract representation multidimensional spaces, and

2. tasks computed on the bases of the positions of entities embedded in these representation spaces.

These types of systems are widely used and are at the core of several industries and research programs, as they have proven to be accurate enough in numerous tasks, as well as popular due to their flexibility. This type of systems includes many of the procedures identified in the machine learning literature, including most recent neural network architectures, specially those known as transformers, and are widely used in social media, but also other parts of the digital public space and in user-facing internet services.

2.2 CONCEPTUALIZING THE IMPACT OF AI IN DEMOCRACY

Conceptualizing the impacts of AI on democracy calls for a second exercise of conceptual delimitation. Instead of conducting an examination of the long list of different definitions of democracy that have been put forward in democratic theory, this article will draw elements from several propositions (Przeworski 2018; Dahl 2020). As such, the conclusions of this article do not hinge on the adoption of any particular definition of democracy. Instead, we consider democracy to be a multifaceted concept and consider elements from several definitions to highlight different areas of impact of AI. This choice is preferred to move to the exposition of the points of contact between the properties or conditions for democracy with politics, and specially in regards to the functioning of digital public spaces such as social media platforms, and the algorithms that mediate those spaces. For an exposition on how different aspects of AI systems have an impact in an exhaustive range of functional elements of democracy, the reader is referred to the work of Jungherr (2023).

> **Among the numerous points of contact between AI and democracy, those in which the impact of AI is conceptualized through their mediation of the digital public space have come to the attention of researchers because of the parallelism between the adoption of social media and a degradation of the quality of the democracy around the world.**

This degradation or democratic backsliding (Hyde 2020) is the object of an ongoing debate on the multi-faceted aspects of this phenomenon. Beginning in the early 2010s, most world regions have seen a decline in the mean value of the Liberal Democracy Index (LDI), as well as an increase in the number of countries undergoing authocratization, according to the V-Dem Institute (2019) report. The LDI index is composed of two factors: the Electoral Democracy Index (EDI) and Liberal Component Index (LCI). The first (EDI), is a systematic measurement of the presence of the minimal elements required by democracy as articulated by Dahl (1971)[9]. The second (LCI), supplements the criteria of EDI by measuring rule of law, respect for civil liberties, and constraints on the executive branch of government by the judiciary and legislative powers.

9 More methodological details on the measurement of these elements is provided by the work of Teorell et al. (2019).

Upon examination, several of the elements of democracy included in Dahl's definition may be argued to be impacted by both the proliferation of a digital public sphere, and AI algorithms mediating these spaces, specially: **1)** freedom to form and join organization, **2)** freedom of expression, **3)** right to compete for votes, **4)** access to a plurality of information, and **5)** the capacity of individuals to organize and express preferences to affect government policy.

1. Freedom to form and join organizations.

The notion of organization and membership has been profoundly changed by social media, lowering costs of organization and mobilization, but also transforming the notion of membership. The Yellow Vests protests in France in 2018, for instance, were first organized as Facebook groups, with media outlets periodically reporting on the number of individuals affiliated with these groups (Ramaciotti Morales et al. 2022). At the same time, visibility of a group while navigating the social platform is a precondition to the freedom of choosing to join it or not, with visibility being largely determined algorithmically in the computation of recommendations. Even for more restrictive definitions of organizations, for individuals accessing information mostly online, visibility of organizations they can potentially join hinges on recommendation algorithms. The impact of AI on this element of democracy is conmesurate of course with the degree to individuals become aware of organization via the internet.

2. Freedom of expression.

While many social platforms lower barriers for information creation and dissemination, freedom of expression on these platforms depends nonetheless on the ability of messages to be presented to the attention of the public. Twitter, for instance, because of the particular affordances on which the platform is built, enables any user to produce content on a par with those produced by political figures, public personalities, or established media. However, most posts on the platform are rarely read or receive scarce engagement (Gabielkov et al. 2016), amounting to absence in the collective deliberation. Intermediate situations resulting from the disparate level of visibility of actors online, and resulting mainly from the algorithmic choices (*i.e.*, whose post is shown in feeds or walls), highlights the inherent lack of equal status in public debate (Habermas 2015). While, of course, public debate unfolding on platforms is not the whole of the public sphere, there is an impact of AI that is commensurate with the level of adoption of social platforms among the public, and specially when online debates and exchanges are sources chosen by journalists, and where a

sizable number of individuals access them for information on the current agenda. A different form of impact relates to the use of AI directly in the algorithmic moderation[10] of posts in view of exclusion from the platform on the basis of the legality of their content and compliance with the terms of use of platforms (Gorwa et al. 2020).

3. Right to compete for votes.

It follows that—to the degree that citizens form their political opinions, beliefs, and preferences based on online content—algorithmic amplification or dimming of messages produced by contenders in elections (in the form of recommendations) may play a role in elections. Social platforms have been shown to be extensively used in elections settings by candidates (Jungherr 2016). Additionally, the use of AI may disrupt this competition by assisting the creation and distribution of content during elections (see recent examples of Russian AI assisted propaganda efforts on Twitter during the 2016 elections in the US, Bail et al. 2020, or in Facebook and Instagram during the EU parliamentary elections; Bouchaud 2024), or by further personalizing the delivery of political messages (*e.g.*, as in Cambridge Analytica scandal).

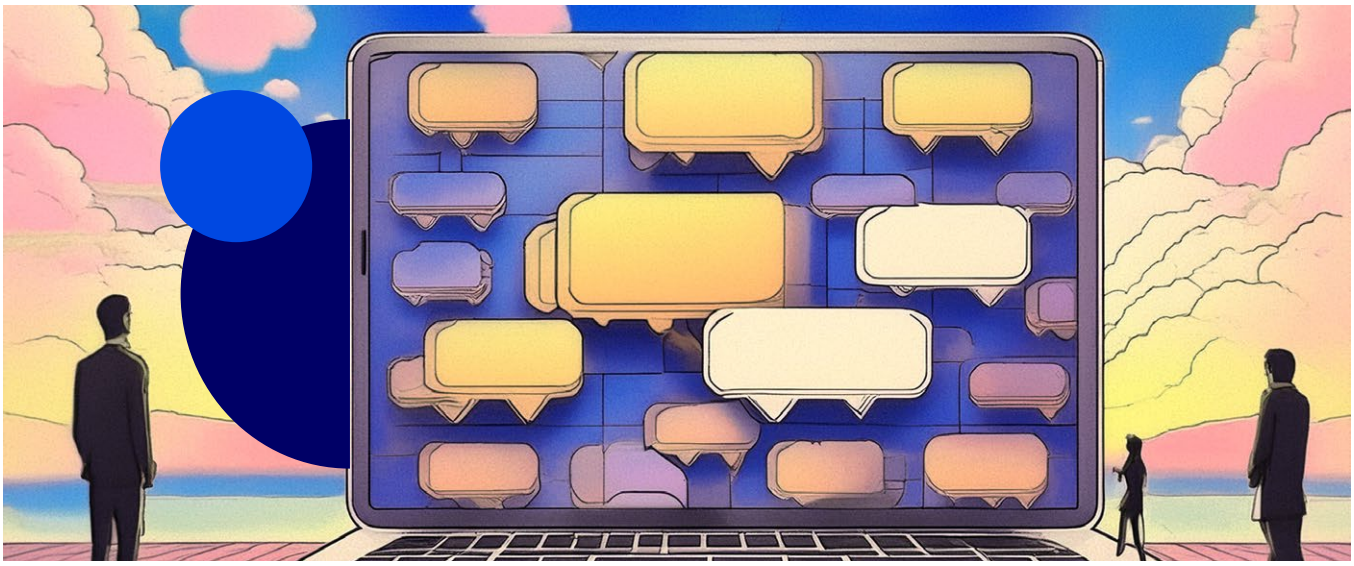
4. Access to a plurality of information.

Commensurate with the fraction of the content that is obtained by individuals through AI-mediated platforms, selective algorithmic exposure plays an important role on the diversity of content in media diets (Kulshrestha et al. 2015). A crucial element of analysis in large social media studies is the measurement of the political diversity of contents proposed to and consumed by users (Bakshy et al. 2015).

5. Capacity to affect government policy.

Whenever a part of the capacity of the public to affect government policy relies on online coordination, AI will be a determinant of this capacity too. A notable case of this is provided by the MeToo movement. Born from a Twitter hashtag (#MeToo), and thus dependent on both the platform affordances and its AI recommendation system (trending hashtags presented to the attention of users depend on algorithmic recommendation), the impact of this collective movement can be shown to have had impacts in different policies (for instance, in the eponymous "#MeToo Bill"; 115th Congress of the US 2017).

10 One prominent example of this form of algorithmic moderation is the use of algorithms to assist hate speech identification in platforms (MacAvaney et al. 2019).



These previous examples do not depend on the volume of information available in a platform: *e.g.*, at a given moment in time, with a fixed number of users and posts, these examples apply in showing different points of contact between AI and elements considered in several definitions of democracy. Additionally, however, it must be considered how AI changes the dynamics of the populations of users (*e.g.*, by facilitating the presence of bots) and that of the pool of posts or messages in platforms (*e.g.*, by facilitating the automation of a part of the effort required in content creation).

In summary, the far reaching implications of computation in society and AI mediation of the online sphere can be structured along some of the elements integrated in several conceptualizations of democracy, specially through their role in politics and, more broadly, in the public sphere.

However, in conceptualizing theoretical frameworks on which to evaluate and improve AI systems, the question remains as to what political structure to leverage in evaluating plurality or visibility. An ubiquitous form of operationalization, rooted in US politics, lies in binary Democrat-Republican ontologies (Adamic and Glance 2005) or single-dimensional analyses (Guess et al. 2023) along a Liberal-Conservative political dimension, on which plurality and visibility are measured. These operationalizations are often deployed in studies seeking to connect AI mediation, and in particular their impact in political segregation, with polarization or other effects in democracy, often pertaining to one of the elements here identified. In the next sections, this article will develop in detail the adequacy of such operationalizations and its relation with the analysis of AI, and specially representation learning systems.

2.3 AI MEDIATING THE DIGITAL PUBLIC SPACE

The unfolding of these profound changes operated to the functioning of public sphere[11], highlighted by the elements of democracy they impact, were initially accompanied by a number of optimist outlooks.

> **The democratization of information creation was to unleash new forces capable of institutional transformation by changing the logic of social mobilization.** (SHIRKY 2009).

Several works link decreasing costs of communication and enhanced tools for digital assembly and collective deliberation with a variety of examples, including for instance the Arab Spring (Lim 2012; Tufekci and Wilson 2012) or the Black Lives Matter movements (Gallagher et al. 2018). A number of voices have nuanced, however, these expectations, highlighting how offline structures reproduce online (Morozov 2011)[12]. At the onset of the emergence of social media, a number of critiques were also formulated regarding how these digital environments and their algorithmic curation could exacerbate personalization and thus segregation (or "cyberbalkanization") and polarization (Sunstein 2001), fostering a range of online "disorders" (Benkler et al. 2018). While some level of polarization is an inherent feature of democracy, the logic spelled by this line of narratives points to the possibility of acute lack of plurality in information, high levels of segregation, and the possibility that they may foster pernicious polarization (McCoy and Somer 2019).

These concerns have driven a large and growing body of scientific literature studying disorders of these digital public spaces, and that have popularized new terms such as *echo chambers* (the relational segregation of online communities along political lines) or *filter bubbles* (a state of affairs in which such segregation is fostered or sustained by the relational configuration created by algorithmic recommendation in social platforms). This latter narrative in which segregation is hypothesized to be caused—at least in part—by algorithms is also motivated by the economic logic underlying algorithmic recommendation in social platforms. Accessible mostly without monetary cost for users, platforms rely on advertisements viewed by individuals during navigation or scrolling, which is

11 The reader is referred to the work of Jungherr (2023) for a list of additional points of articulation between the public sphere and AI systems.

12 A concrete example can be found in the phenomenon of reproduction of offline class structure in digital platforms, determining social gaps and inequalities in the opportunities to create online content (Schradie 2011).

maximized with the time spent on each session. The consequences of algorithmic curation that maximizes time spent on platforms have raised the question of their consequences in terms of political opinion dynamics, polarization, and fragmentation of the public sphere into groups. Some empirical results, for instance, have linked the maximization of engagement in platforms with algorithmic propelling of contents that might exacerbate polarization (Chavalarias et al. 2024). The emerging research, however, paints an heterogeneous landscape. While some cases have been reported of algorithms reinforcing segregation in online content consumption (Roth et al. 2020), several studies have also reported a related increase in the diversity of contents (Aiello and Barbieri 2017). A recent systematic review of causal mechanisms in social media, found that the number of studies suggesting that platforms increase the diversity in content consumption nearly doubles the number of studies that suggest the opposite (Lorenz-Spreen et al. 2023). This family of concerns is reflected in the number and tone of press articles about this matter, in the volume of scientific works dedicated to analyzing the impact of algorithms on politics and democracy , and in the recent regulation specifically targeting political ads online. The Digital Services Act of the European Union, for instance, includes provisions forbidding the computation of recommendations based on political profiling (Digital Services Act 2022, Article 26) and ensuring pluralism online (Digital Services Act 2022, Article 34).

> **The design of algorithms that might minimize or avoid negative impacts related to personalization is an active research domain spanning for at least two decades** (ZIEGLER ET AL. 2005)[13].

This literature is, however, concerned with a particular formulation of this problem traditionally cast as a compromise arising when maximizing both accuracy and diversity of recommended contents (Zhou et al. 2010). This article claims that the prevalence of AI systems that rely on representation learning provides an opportunity to formalize this problem in a new light, explicitly identifying, measuring, and constraining the knowledge that an AI system might have formed about the political configuration of the system that produces the data with which it is trained. This new formulation builds on the one hand on spatial representations of political systems, and on spatial representation of input data made by AI systems on the other.

13 The reader is referred to the work Bobadilla et al. (2013) for a survey on the taxonomy of recommendation procedures and the evaluation of the quality of recommendations with regards to the diversity of contents.

3. THE POLITICAL EMBEDDEDNESS OF THE PUBLIC SPHERE AND THE GEOMETRY OF POLITICS

Different actors are concerned with the role of AI in the digital public sphere, the most relevant for the exposition of this article being those in scientific research, tech industry, and regulation and policy communities. A main concern in these communities is the measurement of the impacts of AI. Specially for platforms, it is important to develop tools for measuring impact in view of public scrutiny, for proving compliance with regulations[14], but also in the development of tools for improving design of algorithms. Addressing the question of measurement raises the need for a framework on which to give meaning and operationalization to the notions of visibility and plurality of content, at the core of the phenomenon of selective exposure, and thus central to political segregation. A traditional framework leveraged in the study of political content diversity in online platforms is provided by political opinion spaces.

3.1 SPATIAL MODELS OF POLITICS

Spatial models in politics seek to explain empirical observations using geometrical features, such as distance and order. In political spatial models, entities are conceptualized as having positions in a (potentially multidimensional) political opinion space. Entities are defined by the ontology of the study at hand, and may include political figures, political parties, individuals of the public, institutions, and even content or pieces of legislation. The dimensions spanning political spaces encode positive or negative attitudes towards issues of relevance for a study: *e.g.*, dimensions measuring attitudes towards income redistribution, or the European integration process, ranging from values 0 for most opposed, to 10 for most favorable (Jolly et al. 2022). The usefulness of spatial models hinge on their ability to link empirical observations with geometrical features: *e.g.*, linking distances between candidates and voters with propensity to vote (Lewis and King 1999). A crucial methodological and substantive challenge in spatial models for politics is the determination of the set of dimensions and the estimation of the spatial position of the entities in a study.

14 The Digital Services Act of the EU, for instance, mandates that large social platforms must produce annual reports of risk assessment, including questions information plurality in the platform, among others (Digital Services Act 2022, Article 15).

DIMENSIONS.

When conceptualizing social systems structured along politics dimensions, the first distinction to be made regards the notion of cleavages (Lipset and Rokkan 1967): enduring and structuring political divisions underpinned by functionalist logic (Parsons et al. 1953), separating, *e.g.*, owners and workers or church and state. Cleavages entail socio-structural divisions, normative elements as values and beliefs, social identities with respect to groups participating in political divides, as well as interactions, institutions, and parties (Bartolini and Mair 2007). A central and challenging methodological and conceptual task in determining such lines of divide and the issues dimensions associated with them is the measurement of the social structuration of political composition (Marks et al. 2022). The consideration of this family of political dimensions stems from political sociology.

A second type of consideration in identifying political dimensions stems from strategic and game-theoretical dynamics proposed in political economy (Downs 1957). In this theoretical setting, actors such as candidates and parties constitute the *supply* side, as they offer political representation with given positions on issues. Other actors (typically voters, but more broadly the public) in the *demand* side occupy positions on relevant issue dimensions in accordance to the utility extracted from them (*e.g.*, adopting positions on the issue of redistribution and taxation depending on income), and choose political offers on the basis of their positions. In this theoretical framework positions of actors on multidimensional issue spaces obey not only a social logic of structuration, but also strategic considerations of political gains and utility extracted by adopting and adapting their positions (Riker and Ordeshook 1968).

In addition to issue dimensions, one tradition identifies ideological dimensions as indicators of positions along several *aligned* issues (Converse 1964; Jost et al. 2022). For instance, in the United States, an ideological Liberal-Conservative dimension is often considered in social media studies. The position of individuals along such a scale is informative of several issues at once (*e.g.*, abortion, gun control), provided that they display high alignment: *i.e.*, that the position on one is constrained or dependent, and thus informative, on the other in a descriptive settings. Several studies focused on U.S. settings leverage a unique Liberal-Conservative dimension due to the traditional high alignment of relevant issues[15]. In general, however,

15 Some scholars have proposed that different moments and spheres in the U.S. might be best represented by an additional and independent (*i.e.*, not aligned) political dimension. See the work of Uscinski et al. (2021) for a recent example.

it is not possible to reduce social systems to a single dimension when seeking to explain empirical observations with spatial models. European countries, for instance, display an heterogeneous dimensionality (*i.e.*, the number of dimensions needed to account for observations; Benoit and Laver 2012). **Figure 1** (from Ramaciotti Morales and Vagena 2022) illustrates this in a spatial representation of political parties in European countries along different dimensions, as measured using expert surveys: surveys administered among experts in political sciences and in which they position parties along several issue and ideology dimensions.

MEASURING POSITIONS ALONG DIMENSIONS.

Expecting actionable low dimensional representation in many settings[16], the use of spatial models has been fruitfully leveraged in very wide scope of studies. A crucial aspect of these representations is their ability to enable and determine observable actions. Among these, an important type of cognitive representation is that of evaluative opinion, which allows for individual dispositions *for* or *against* something (Bem 1970), operationalized as attitude: an individual disposition towards an attitude object (*e.g.*, person, institution, issue, event), which is determinant of behavior (Ajzen 1989). Attitudes offer a conceptual link between internal representations and external observable behavior through their evaluative function (with examples ranging from expression of cultural tastes, Sonnett 2004, to political donations, Bonica 2014, or voting, Gerber and Lewis 2004), in a way that is operationalizable via spaces. In these spaces dimensions stand for attitude indicators, providing numerical quantification of positive or negative stances towards different attitude objects. Observed behavior (*e.g.*, uttered opinion, vote, elicited response) may not always match internal attitudes (*e.g.*, in planned behavior; Ajzen 1985).

Measuring positions of actors on issue or ideology dimensions has been traditionally approached from two methodological perspectives. The first one is survey research: eliciting self-positioning from individuals or asking experts to position entities such as parties or media outlets. An example of the former are traditional surveys (*e.g.*, asking respondents to position themselves on issues in numerical scales).

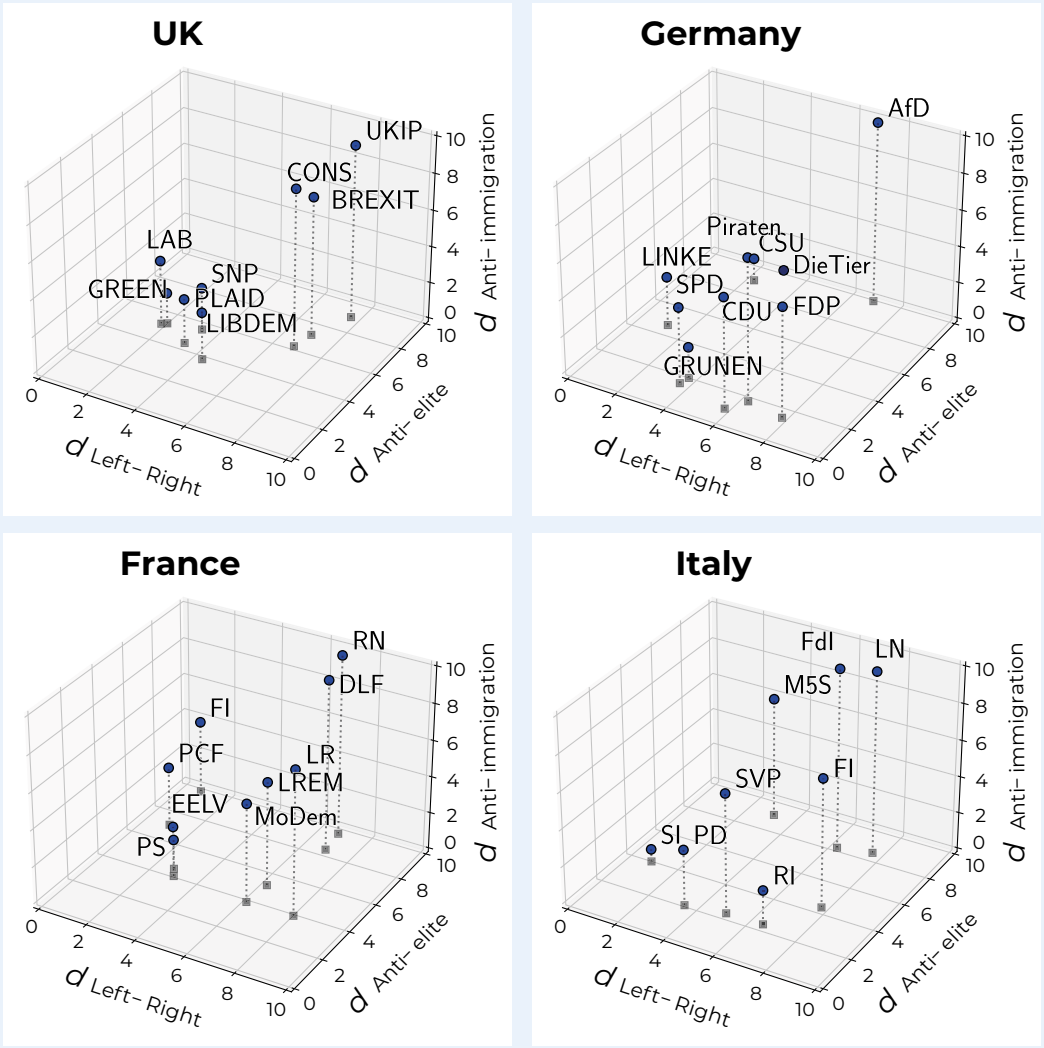
16 The ability of individuals to distinguish and organize their environments depends on their capacity to make meaningful mental representations with bounded cognitive resources, selecting only a few dimensions or organizing them in bundles reducing the dimensional complexity of the perceived environment (Converse 1964; Hix 1999). Similarly, because of institutional and organizational constraints, but also because of the bounded cognitive capacities of members and candidates, political groups may also produce an offer of limited dimensional complexity to be appraised more easily or positively by voters (Olbrich and Banisch 2021).

An example of the latter are expert party surveys, in which experts in politics are asked to position political parties on numerical scales. A concrete illustration is provided by the Chapel Hill Expert Survey, in which hundreds of experts position European parties on tens of issue and ideology numerical scales (*e.g.*, redistribution, European integration, liberalization of immigration) ranging from 0 (most opposed) to 10 (most favorable). A second methodological approach to measuring positions of actors is statistical estimation using data traces: behavioral (*e.g.*, clicks, votes, interaction networks) and text data produced by these actors. The inference of positions leveraging spatial representations of political opinions can be traced back at least to the NOMINATE[17] method (Poole and Rosenthal 1985, 2000), conceived to estimate ideological positions of members of parliament in the U.S. using roll call data (*i.e.*, registries of how they vote bills).

FIGURE 1.

Party positions in the UK, Germany, France and Italy along three dimensions (Left-Right, anti-elite sentiments, and stances against liberal immigration policies), computed as the mean positions on a scale from 0 to 10 attributed by experts in the 2019 Chapel Hill Expert Survey (CHES; Jolly et al. 2022).

(Source: Ramaciotti Morales and Vagena 2022)



17 NOMINATE stands for *Nominal Three-Step Estimation*.

Social media data, because of their volume, diversity, reach, and granularity, have given rise to a stream of research on political position estimation.

(SEE MESSAOUDI ET AL. 2022, OR LIU AND ZHANG 2012, FOR A RECENT EXTENSIVE REVIEW).

Inferring opinions of users in social media platforms is crucial not only in the investigation of algorithmic mediation and selective exposure (*e.g.*, the effects algorithmic recommendation; Ramaciotti Morales and Cointet 2021), but also in investigating broader social and political phenomena (*e.g.*, social mobilization, Budak and Watts 2015, Cointet et al. 2021, agenda setting dynamics, Barberá et al. 2019). The field of research interested in the use of trace data for the inference of positions in geometrical spaces where dimensions indicate positions or ideologies is traditionally referred to as *ideology scaling* or *ideal point estimation* (Clinton et al. 2004). Similar methods have been used in political science to compute ideological positions of donors, judges, and in general any actor capable of producing observable choice behavior (see Imai et al. 2016 for a survey, and Peress 2022, for a recent example of an ideology scaling method).

3.2 THE GEOMETRY OF ONLINE POLITICS

The statistical inference of issue and ideology positions using trace data produced in online social platforms is a recent but very active field of research. The results from this line of research are central to the study of algorithmic mediation and selective exposure, as they allow to connect the political position of users and contents (as conceptualized in multidimensional political spaces), with recommendations produced by the platform, as well as with actions taken by users (*e.g.*, sharing or commenting content to which they've been exposed). One of the first studies to pioneer ideology scaling from social media data was that of Barberá (2015). In this study, Barberá used Markov Chain Monte Carlo methods to infer the ideological position of millions of U.S. Twitter users on a Liberal-Conservative scale using observed follower (*i.e., who follows whom*) networks on the platform, leveraging underlying psychological mechanisms such as homophily (in particular *value homophily*; Lazarsfeld et al. 1954) through Bayesian inference. Similar frameworks have been proposed in a multitude of settings, and using diverse behavioral data traces. Bond and Messing (2015) used likes given on Facebook as behavioral traces to infer stances of a large number of users in a similar way. Behavioral data traces are determined by the specific affordances of each platform, with observable actions including, *e.g., follow, share, like, subscribe,* or *upvote*, counting heterogeneous possibilities across platforms. These interactional traces linking users and items (*e.g.*, a post, a piece of content, a URL), or users to other users, form a network in the sense that they link two entities. These methods do not necessarily rely on textual data, making them text-independent and thus also language-independent, which makes them especially valuable for analyzing cross-national settings[18] (Barberá 2015; Ramaciotti Morales and Vagena 2022).

A different and popular family of methods for political position estimation uses textual traces. The first methods for ideological positioning of texts draw inspiration from interactional methods: texts may be ideologically positioned according to whether they included or not particular keywords that might be indicative of ideology (*e.g.*, the Wordscore method of Laver et al. 2003, later adapted to Bayesian frameworks; Slapin and Proksch 2008). More recent approaches for text-based inference include Markov Chain Monte Carlo Methods, and more recently variational inference (Vafa et al. 2020). A growing number of works is proposing the use of unsupervised spatializations of text and documents (and actors producing texts)

18 Regarding platform regulation, this is specially relevant for regulatory settings involving multi-lingual bodies such as the EU.

using text embedding methods (Mikolov et al. 2013; for an extensive review see work of Jurafsky and Martin 2022), or Large Language Models (LLMs) in identifying ideological positions from text (Wu et al. 2023).

Either on the basis of text or other data traces, this set of works shows that it is possible to represent online populations and online contents in political opinion spaces.

A concrete illustration of this process, applied to multidimensional European settings, is provided by a method called Language-Independent Network Attitudinal Embedding (LINATE; Ramaciotti Morales and Vagena 2022). As its name suggests, it leverages interactional data traces and is thus language independent. The method works as in two phases (see Ramaciotti Morales et al. 2022 for additional details). In the first phase, interactional data traces subtended by an online population (*i.e.*, network data) are used to produce a first multidimensional embedding. In the second phase, the multidimensional positions of the individuals of this population are mapped onto a multidimensional *referential* political space. In the study of this example (Ramaciotti Morales et al. 2022), the first phase is conducted on a Twitter population selected among users involved in the French Twitter sphere, subtending a large follower network. The second phase is computed using the above-mentioned Chapel Hill Expert Survey (CHES) data as a referential political space. The map between the first embedding and the CHES space composed of 51 dimensions, is computed using as reference points the positions of political parties: these positions are readily by provided in the CHES data, and in the first embedding space they are computed as the mean position Members of Parliament (MPs) present on Twitter and that belong to each party.

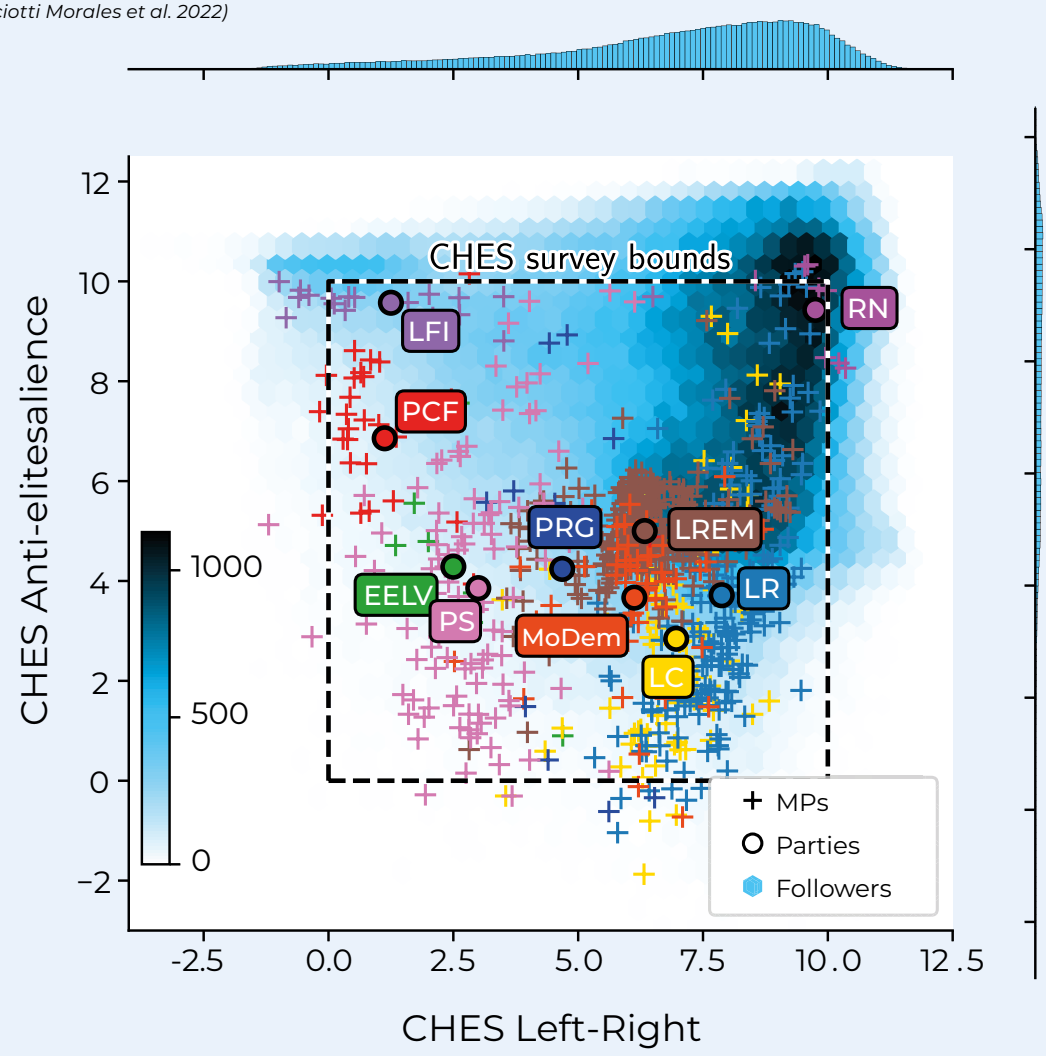
Figure 2 (from the study reported in Ramaciotti Morales et al. 2022) shows the spatial distribution of the large population of study (230.911 Twitter users) along the two irreducible (*i.e.*, non aligned) dimensions that are most explicative of the network data observation: a Left-Right dimension, and a dimension measuring attitudes towards elites and institutions. **Figure 2** also shows MPs colored by party and the position of parties as provided in the CHES data. Importantly, these dimensions are those of the CHES survey instrument, and comme endowed with reference points: positions 0 and 10 mark the leftmost and rightmost positions for political parties on the Left-Right scale, while positions 0 and 10 on the anti-elite dimension mark respectively the positions having the least the most prominent anti-elite sentiment. In the next

section, we will use this population to illustrate how multidimensional political positions can be leverage in AI explainability with regards to politics, *i.e.*, in measuring the the amount of political information learned by an AI system that has access to and is train on the data produced by this Twitter population. It must be noted that, while following this particular spatialization method as illustration, any method providing a political spatialization of online populations can be leveraged in AI explainability. It must also be noted that political spatialization concerns users, but also others entities, such as parties (as seen in **Figure 2**), but also news contents (Ramaciotti Morales et al. 2023), web domains (Cointet et al. 2021), YouTube channels and Facebook groups (Ramaciotti Morales et al. 2021), reaching to potentially any entity to which interactional or text data can be attached.

FIGURE 2.

Example of a multidimensional spatialization of online populations (users of the French Twittersphere), positioning entities (in this case users) along two irreducible political dimensions relevant in France. The density of positions is shown in shades of blue. The positions of MPs are shown in colors by party. Spatial positions are calibrated using the Chapel Hill Expert Survey data.

(Source: Ramaciotti Morales et al. 2022)



4. CAN AI SYSTEMS (INADVERTENTLY) LEARN POLITICAL REPRESENTATIONS?

We now turn to how AI systems produce spatial representations of input data (seen during training, testing or deployment), and specially data produced in online spaces such as social platforms. For the family of AI systems of interest for this article (*i.e.*, those based on spatial representations), providing frameworks of analysis of these representations and examining the role they play in producing recommendations falls within the realm of AI explainability.

4.1 EXPLAINABILITY AND SECURITY

Explanability of AI, as a research field, is not new[19]. The term *explainability* is often used in connection with *interpretability* or *intelligibility*. It is not within the objectives of this article to provide an exhaustive review of the vast scientific field of AI explainability, nor to present a detailed account of the debate regarding the differences between explainability, interpretability or intelligibility. The reader is referred to the work of Adadi and Berrada (2018) for the former, and to that of Marcinkevičs and Vogt (2020) for the latter. Broadly defined, the goal of AI explainability is to improve human intelligibility of computations processes, or to make opaque computation processes understandable. This need originates in the fact that the models on which modern computation processes rely are complex (in size and structure) and that, because of the role that computation has gained in society, important properties linked to human oversight (such as accountability) depend on the ability of humans to understand them (Castelvecchi 2016).

> **The complexity of AI systems stems in part from the fact that computers can treat increasingly large volumes of data with increasingly sophisticated and flexible models.**

Very large physical simulations, while potentially processing more computations than a human could process (or any team of humans for that matter), are structured along intelligible models, resulting in values of parameters and variables that, while hard to compute, are

19 Early examples of systematic explanations of algorithmic decisions can be traced back at least to the 1970s (Scott et al. 1977).

easy to interpret. Recent AI systems, in contrast, provide means to flexibly learn during training a large variety of models. This is the case, for instance, of neural network AI systems, which can emulate linear regressions (Marquez et al. 1991) or classification trees (Bondarenko et al. 2017) among many other models.

Among the several AI explainability taxonomies that exist in the literature, we illustrate their scope with the example of that proposed by Zhang et al. (2021), distinguishing four families of explanations by increasing explanatory power[20].

1. **Examples.** A first ambition in AI explainability is that of providing examples; *i.e.*, assembling chosen prototypes of stable inputs and outputs that illustrate the functioning of a system. If we consider, for instance, an image classifier AI system, this amounts to a collection of images with the labels that the system consistently outputs for them.

2. **Attribution.** A second ambition, called attribution, is that of providing a measure of the impact of some of the features of the input data. In our example of an image classifier, this amounts to linking features (mean color, brightness, or more sophisticated properties) to some of the classes that the classifier can propose as output. Have images classified as apples comparatively more color red pixels than those classified as pineapples?

3. **Hidden semantics.** A third level of explanation seeks to map attributes of input data to the representations, or hidden layers[21] of an AI system. In our example of an image classification system, an instance of this would be to inspect the representation of the input images made by the system (in an embedding space in a transformer, the columns space in a matrix factorization algorithm, or the weight space of a neural network) to map a feature of these input images to a dimension of the representation space (or different, more complex, geometrical features). When shown an image of an fruit, does the AI system embed it on a representation space in which a dimension seems to order them by amount of red pixels, always putting apples to the right of pineapples?

4. **Explicit rules.** A four level explanation consists of predicating explicit operational rules emulating the behavior of the AI system. In our example of image classification, an instance of this level of

20 This taxonomy was proposed as a result of a survey on the interpretability of neural network AI systems. Thus it targets systems that are by design opaque because of the flexibility of models achievable by Neural Networks.

21 Hidden layers are ubiquitous features in several AI systems, also called black box, for its opacity. Mathematically, they may take the form of weight spaces in Neural Networks, but a diversity of forms exist depending on the AI model, including column space in matrix factorization systems, or embedding in transformers.

explanations would be to establish an intelligible operational model that predicts the output of the image classifier. For instance, establishing a logistic or polynomial regression model[22] depending on features of inputs images, and that emulates the behavior of the AI classification system.

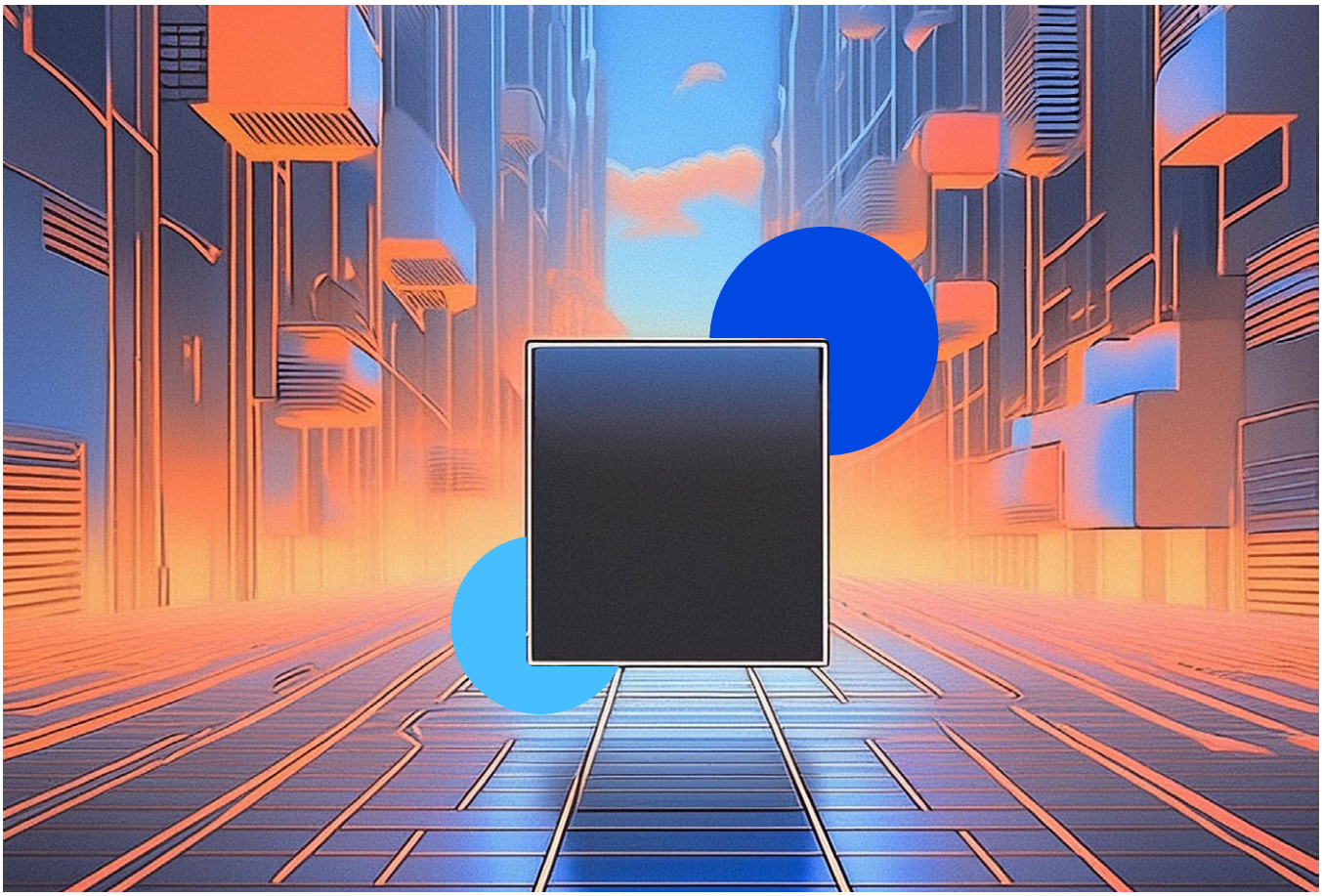
Other surveys on AI explainability propose different taxonomies, but most include a level of explanatory power centered in the ability to attribute semantics to hidden representations (see for instance the survey by Adadi and Berrada 2018). In increasing order of explanatory power, hidden semantics is the first type of explanations of interest for this article.

Whenever the computation of an AI system can be mapped to operations taking place on a spatial representation for which intelligible semantics are available, it is possible to cast a large family of AI security problems in geometrical terms.

Applied to the case of political segregation in online platforms, this type of explanations enables the design of algorithms improving risk management linked to lack of diversity in content recommendation and polarization. This connection will be made explicitly in the next sections, but briefly summarized, the main claim is the following. Instead of prescribing a level political diversity of recommendations, once a dimension of the machine representation has been associated to a political dimension (*e.g.*, a Left-Right dimension), this knowledge can be leveraged in modifying the representation space to exclude such dimensions from downstream applications such as recommendation.

This path connecting explainability with new design tools for moderating risks related to political segregation and polarization incurs two important challenges. First, the task of attributing hidden semantics of political nature to representations learned by AI systems is a challenging one. In other words, if it is deemed that a given set of political issues and ideology dimensions are of importance for a given online population (based on a theoretical and empirical political science framework, such as in the example from the previous section), it is often challenging to link a spatial dimension of a representation learning space with a political issue or ideology dimension.

22 The key property of this example is intelligibility by humans, as logistic regression is a akin to a generalized linear model because the outcome always depends on the sum of the inputs and parameters.



Let us illustrate this challenge through a concrete example. If Left-Right or Liberal-Conservative are deemed to be political dimensions of importance to which we would want to render an AI system blind by design, it is not trivial to go inside the black box to identify which model dimensions are related to these political dimensions. Secondly, even if a political dimension relevant in structuring an online population was identified to be encoded in the spatial representation created by an AI system, the way in which the political information is encoded in this space might not be readily treatable because of the geometrical complexity with which it is encoded. For instance, if we consider a set of users or contents for which we know their relative order going from Left- to Right-leaning, the way in which they are positioned in the machine representation space might be mediated by a non-linear map; *i.e.*, in representation space, they might be ordered from Left to Right on a geometrical pattern different from a line or spatial direction.

> This problem is framed in the context of machine learning as the linear representation hypothesis.
>
> (MIKOLOV ET AL. 2013)

In the rest of this section we will address the first of these two challenges, leaving the second one for the next sections.

4.2 AI REPRESENTATIONS OF ONLINE POLITICS

This subsection presents the results from the work of Faverjon and Ramaciotti (2023) to illustrate the possibilities and challenges in making a theoretical and operational connection between

1. dimensions of political analysis as conceptualized in political sciences, and
2. dimensions of AI representation spaces.

This study leverages the same population of Twitter users from the French sphere of **Figure 2** from the study by Ramaciotti Morales et al. (2022), spatialized along two irreducible dimensions that are most structuring of that particular ecosystem (Left-Right and anti-elite dimensions). Using behavioral trace data produced by individuals in this population, a recommender system is trained to recommend content from media outlets, following a standard training procedure in online platform settings. The recommender system chosen is a widely used matrix factorization algorithm (Boutsidis and Gallopoulos 2008), in which the representation space computed by the AI system is spanned by the column space of the resulting matrices.

> Having access to political positions of individuals on the one hand, and to the machine spatial representation produced during training on the other, the cited study proposes one alternative to tackle the AI explainability problem: using multidimensional political positions of individuals to attribute meaning (*i.e.*, hidden semantics) to the dimensions of the AI representation space.

The results of this study show that, among the 12 dimensions of the representation space of the chosen recommender system, one dimension can be associated with Left-leaning individuals, and one dimension can be associated with Right-leaning individuals. The rest of this section provides more details on this result to then present how AI explainability based on political dimensions enables the formalization of a new family of problems in AI security tackling the risks of political segregation and polarization.

TRAINING A MEDIA CONTENT RECOMMENDER USING BEHAVIORAL TRACE DATA.

Recommender systems that propose contents in online platforms (*e.g.*, in *feeds* or *walls*) leverage a large and diverse set of signals produced by users. In the cited study, authors show that even a very narrow set of signals may allow an AI system to produce a political representation of the online environment they are mediating with recommendations[23]. For this, the study selects as a signal for training the URLs shared by users in Twitter posts and that link to web content (*e.g.*, articles from established media, independent media, blogs). The task given to the recommender system, and that defines the optimization problem guiding the training, is that of predicting whether a web domain will elicit engagement on the part of a user.

> One important motivation in this research design is the fact that social media platforms are known to act as news and content providers for a large number of users, fulfilling a central function in the online public sphere.
>
> (KWAK ET AL. 2010)

Concretely, from the population of 230.911 Twitter users sampled from the French Twittersphere, and for which political positions are estimated on Left-Right and anti-elite dimensions, a random sample of 40.000 users was selected to then proceed to collect all of the URLs that they have shared online, belonging to 426.014 unique web domains[24]. These signals were then used as data (*i.e.*, user-media pairs) to train the chosen recommender system until it was deemed to produce accurate predictions of user-media pairs[25].

23 This aspect is specially relevant in light of limitations on the processing of data on political opinions, such as under GDPR, and will be discussed in the conclusion.

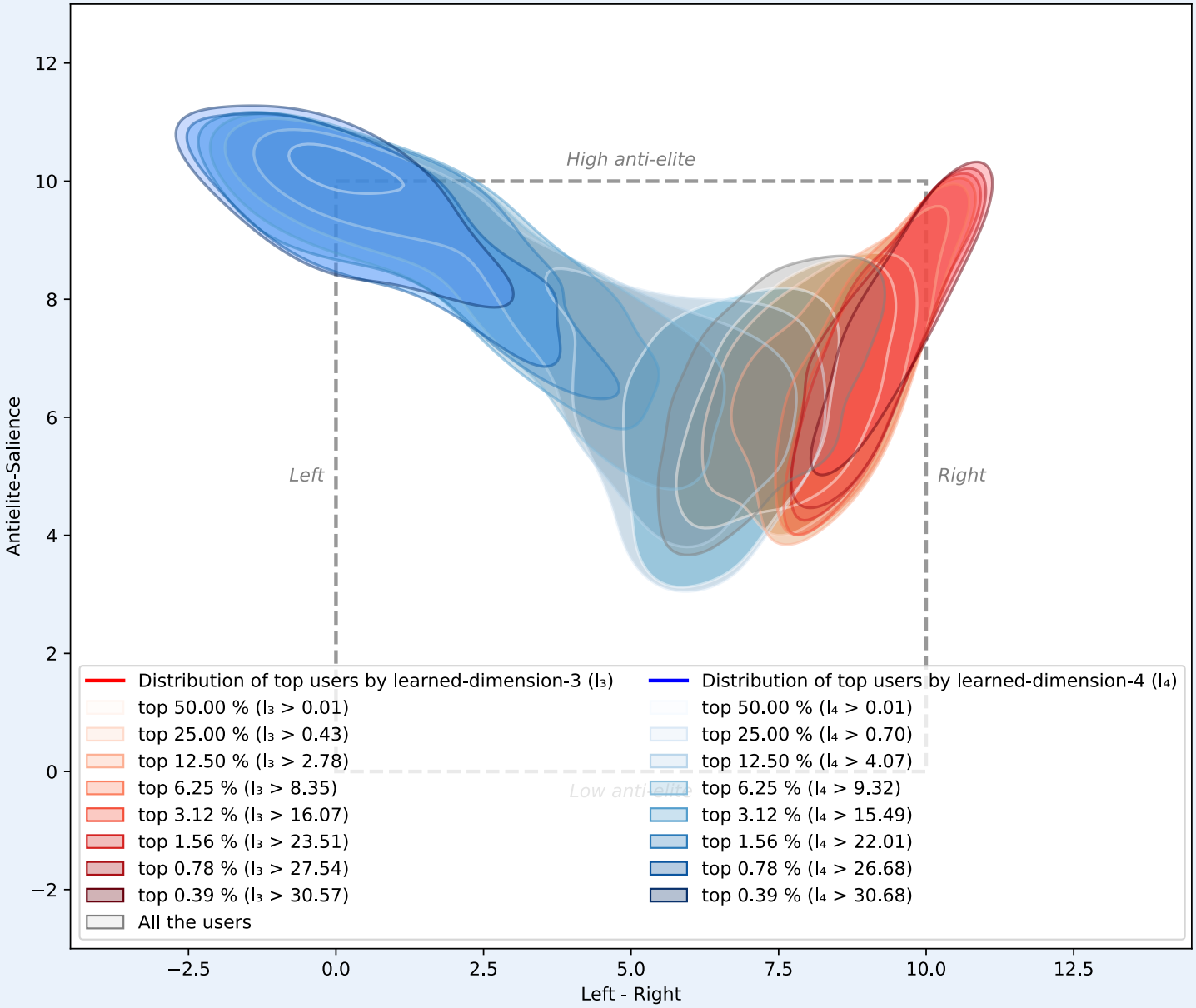
24 A web domain is the root website that may hosts several contents having distinct URLs. For instance, an article from the news media Le Monde in France may have a URL in the form www.lemonde.fr/article, which will be hosted by the domain lemonde.fr. In general, media outlets have unique web domains.

25 During training, a fraction of the user-media pairs were reserved for testing the accuracy of the predictions. Predictions are used in recommendation in the sense that, if the recommender system predicts a user has high probability of taking interest in a media outlet so as to share its contents online, it is probably relevant as content to be recommended.

FIGURE 3.

Spatial distribution of Twitter users on the Left-Right and anti-elite two-dimensional political space according to increasing level of salience in the dimensions l_3 and l_4 (out of 12, l_i for $i = 1, \ldots, 12$) of the AI representation space computed during training in a recommender systems setting. Dimensions l_3 and l_4 are deemed as the most dependent on political positions as quantified by mutual information metrics.

(Source: Faverjon and Ramaciotti 2023)



COMPUTING HIDDEN (POLITICAL) SEMANTICS.

The particular recommender system used in this study generates during training a spatial representation of 12 dimensions in which users and web domains are embedded, with the predicted probability of a user sharing content from a web domain computed on the basis of angular similarity through an inner product. To operationalize the measurement of the amount of information about the political positioning of users that was encoded in each of these 12 dimensions (l_i, for i = 1, … , 12), a mutual information metric was used[26]. The examination of the dependence between the positions of users along political dimensions on the one hand, and machine representation dimensions on the other, revealed that machine dimension l_3 was dependent on the position of users on the Right-wing side of the political spectrum, while dimension l_4 was dependent on the positions of users on the Left-wing side of the political spectrum. This is illustrated in **Figure 3**, showing how users selected by the salience of positions they have along these two dimensions of the machine representation space, project onto regions the political space of **Figure 2** subtended by a Left-Right and an anti-elite dimensions. **Figure 3** shows the spatial distribution of users on the Left-Right and anti-elite two-dimensional political space, by increasing level of positions[27] in the machine dimensions l_3 and l_4. As the positions along l_3 and l_4 increase, the users found in those positions are also shown to concentrate increasingly in the far-Right and far-Left positions respectively.

26 In information theory, the mutual information between two variables is a measure of the dependence between them. In the case of the presented study, mutual information is computed between the political position of users along the two known political dimensions, and the 12 known positions in the AI representation. In other words, mutual information is a quantification of the degree of information that is known about the position of a user in one of the 12 machine dimensions whenever its position is known in one of the political dimensions.

27 In the reported study, the delimitation or regions shown is computed as the level curve of probability equal to0.5 in the Kernel Density Estimation probability function in the political space for a given level of salience l_3 and l_4.

NEW APPROACHES IN AI SECURITY FOR POLITICAL SEGREGATION AND POLARIZATION.

Let us restate the risk of AI mediation in the digital public sphere considered by this article. Because algorithms allows to recommend different content to each user, the possibility exists that the resulting selective exposure to content online is structured along political lines, *e.g.*, recommending only Conservative-leaning content to Conservative-leaning individuals. A central concern associated with this scenario is that political segregation in information consumption may exacerbate political polarization (Sunstein 2001). As a result, there is considerable investment in research measuring political diversity of algorithmic exposures (Bakshy et al. 2015) and in diversification, which incurs a normative problem: How much diversity must be achieved, or, to what degree additional diversity must be proposed during recommendations? This normative approach results in two forms of guidelines:

1. Optimize a utility function during training that integrates both accuracy and diversity (by prescribing their relative importance), or
2. Setting constraints in minimal content diversity that algorithms should propose and then optimize for accuracy.

These approaches incur traditional challenges of normative approaches:

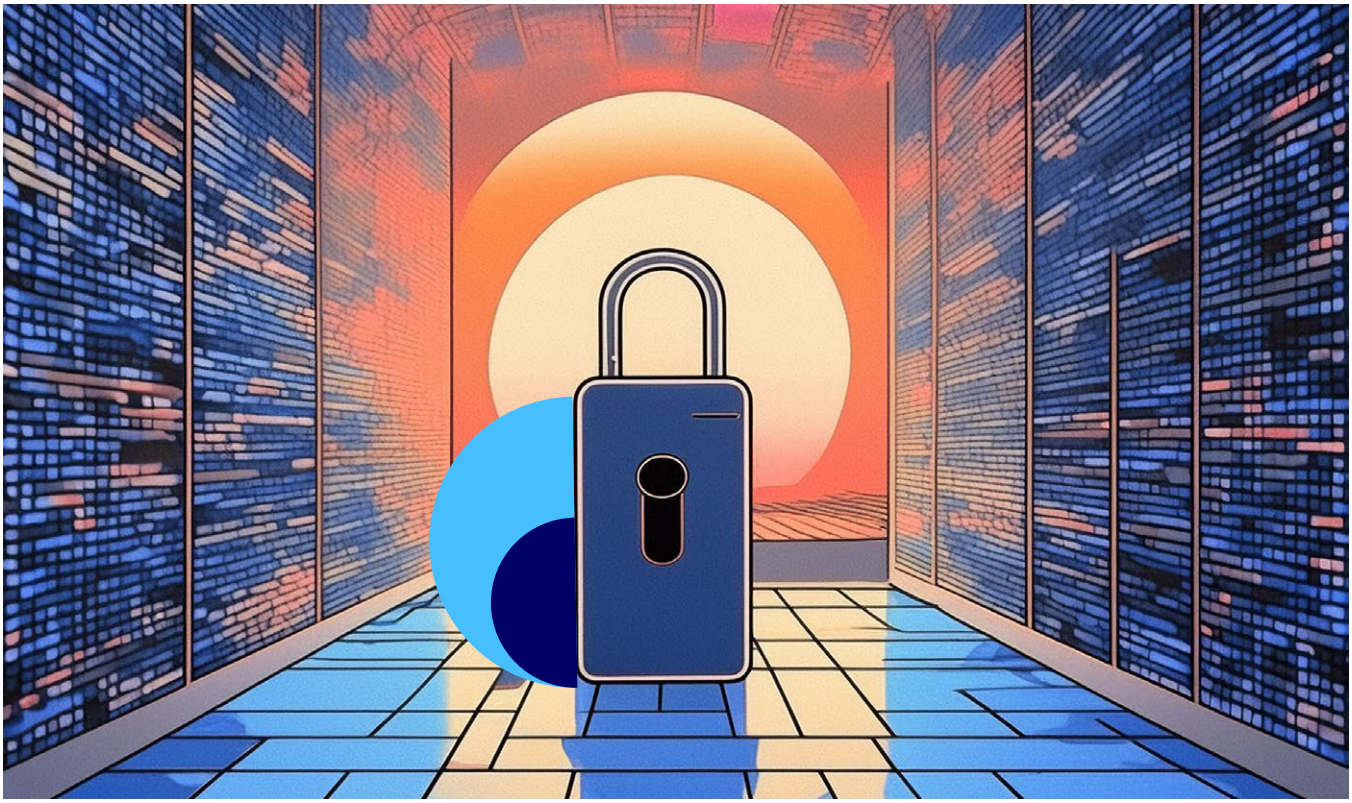
- Who decides the level of diversity to be enforced?
- Who decides on the ontologies on which diversity is to be measured and enforced?
- Should we enforce diversity with regards to Liberal-Conservative divides, but also with regards to a separate issue and which one?

A recent and illustrative debate involving normative diversity in AI (although not completely related to politics) is that followed by the design decision to diversify ethnic representations in images provided by Google Gemini text-to-image generator (Gautam et al. 2024). The case of political opinions or beliefs presents an additional risk. Contrary to several narratives in social media research, recent experimental results in curating diverse political content consumption show that increased diversity may sometimes lead to exacerbated political polarization (Bail et al. 2018). A claim of this article is that AI explainability that hinges on politics, as tackled in the study by Faverjon and Ramaciotti (2023), opens a path to new approaches to AI security. First, by making a theoretical connection with comparative politics,

algorithm design communities may assess the ideologies and issues that are structuring to a particular national setting or even to a given digital arena, and on which the question of algorithmic design and consequences should be focused.

Second, whenever the possibility exists of attributing political semantics to hidden spatial representations leveraged by AI, it is also possible to design AI systems and recommendation procedures that address the impact on politics without incurring normative approaches by removing political information encoded by the machine.

Following the example of the case study by Faverjon and Ramaciotti (2023), once a model is trained to compute recommendations, the machine dimensions identified with political ones that are relevant for the particular arena of deployment (*i.e.*, the French Twitter sphere) may be removed from the computation of recommendations, in a procedure akin to rendering algorithms blind to politics[28].



28 It must be noted that stances on political issues might not be independent from other features in a given population (*e.g.*, age, gender, income), and that removing political information from a spatial encoding might also remove other information . The consequences of this possibility for algorithm design will be addressed in the conclusions.

5. DISENTANGLING POLITICS AND RECOMMENDATIONS: A DATA-DRIVEN EXAMPLE

In order to formalize these challenges and the operational aspects of this new family of AI security approaches relating to politics, this section illustrates a concrete application case by leveraging a synthetic population approach. Synthetic populations have two important advantages on which this data-driven example case study builds. First, they alleviate the need for data that is difficult or impossible to obtain. This is essential in our case, because of the impossibility of accessing the models trained by social media platforms. Second, even if we had access to the trained models included in recommender systems in large online platforms, the inherent unobservability of political opinions would render the exercise of analyzing new algorithmic strategies dependent on the quality of the estimation of these opinions. Using synthetic population in which we can precisely prescribe political opinions and other features, separates the problem of illustrating the new family of algorithmic design problems proposed by this article from the problem of estimating political opinions.

5.1 SYNTHETIC POPULATION SETTING

Let us consider a synthetic population made of N = 1.000 individuals where only three features are important for diversity and accuracy in a recommendation setting. Let us further assume that these features are quantifiable in continuous scales or dimensions, and that the first one quantifies negative or positive attitudes towards a particular political issue or ideology[29].

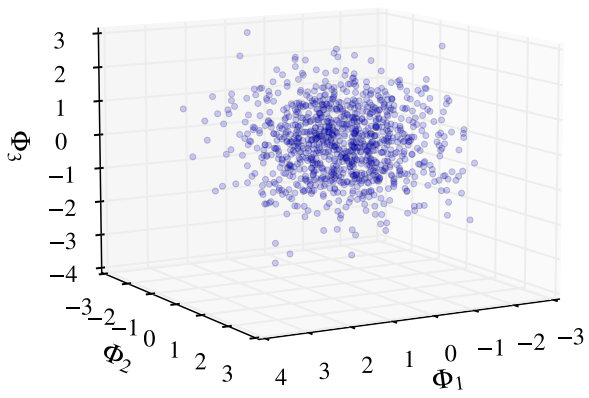
We name these variables Φ_1, Φ_2 and Φ_3 respectively, with Φ_1 representing our political variable, and $\mathbf{\Phi}$ the space subtended by all three dimensions. Additionally, we name the three-dimensional position of an individual i (i = 1, … , 1.000) in space $\mathbf{\Phi}$ as $\boldsymbol{\phi}^i$. Let us prescribe the distribution of values of these 3 features in our synthetic population with a multivariate Gaussian distribution, $\boldsymbol{\phi}^i \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu}$ = (0, 0, 0), $\boldsymbol{\Sigma}$ = diag(1, 1, 1), and draw from this distribution values $\boldsymbol{\phi}^i$ for i = 1, … , 1.000 (see **Figure 4a**).

29 If it is an issue, negative and positive values encode degrees of negative and positive attitudes. If it is an ideology in the descriptive sense, values encode proximity towards opposed ideological positions, such as negative values encoding degrees of Liberal attitudes and positive values encoding degrees of Conservative attitudes, as in most social media studies in US settings (Bakshy et al. 2015).

FIGURE 4.

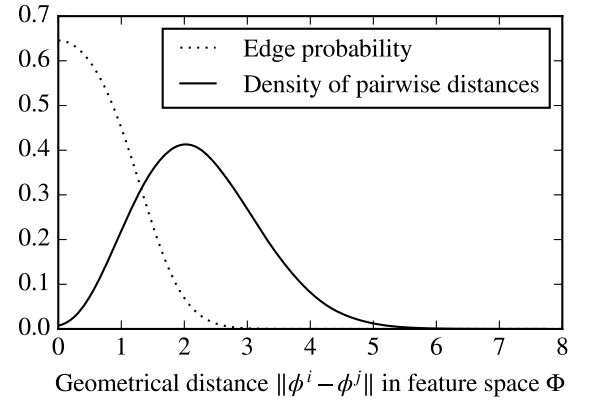
Synthetic population for the recommender system setting.

4a.



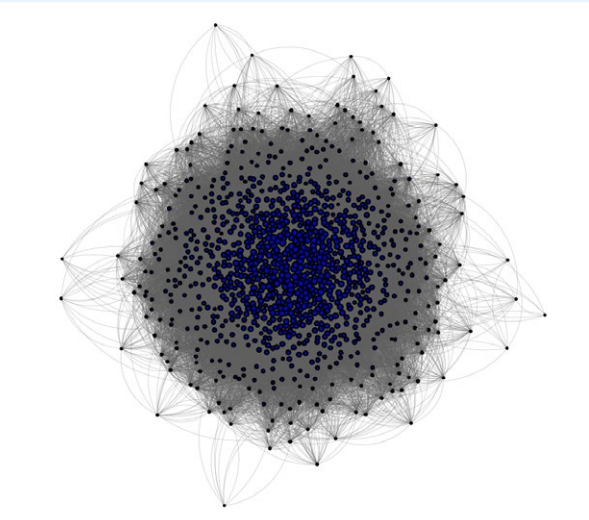
4a. Positions along three feature dimensions for a population of N = 1.000 individuals, drawn from a multivariate Guassian distribution.a multivariate Guassian distribution.

4b.



4b. Probabilistic law prescribing the process for observable interaction data.

4c.



4c. Synthetic interaction network observable by the recommender system during training.

AI systems cannot directly observe all features of individuals (*e.g.*, political opinions), and they are trained instead on the data traces they produce. Let us further specify a data generation process for this population in the form of an interaction network. Concrete examples of interaction networks are follower of friendship networks, but also interactional data indicating who has, *e.g.*, shared, commented or liked content produced by other users, and take the form of relational data that may be represented as edges or links in a network. We prescribe a data generation process based on homophily (Lazarsfeld et al. 1954), one of the most documented, understood, and ubiquitous social processes in social networks[30]. In an homophilic data generation process, the probability of observing an interaction between two individuals (i and j) depends on their similarity, operationalized in our settings as their distance in the feature space $\mathbf{\Phi}$. Formally, this process is prescribed as the probability law for observing an interaction between i and j based on their distance in feature space, which we set as a logistic function of the distance between users i and j in feature space $\mathbf{\Phi}$. This particular form is rooted in Item-Response Theory and is used for political position inference in Bayesian settings in online studies[31], using the observable interactional data and the probability law to estimate values of $\boldsymbol{\phi i}$ (Ramaciotti Morales et al. 2021; Barberá et al. 2015).

Figure 4b shows the density of pairwise distances for our synthetic population in space $\mathbf{\Phi}$ and the prescribed probability of interactions[32]. Using this probability of interactions, we draw edges to constitute our synthetic interaction network of data traces that will be observable by our recommender

30 The reader is referred to the work of McPherson et al. (2001) for an extensive survey on the role of homophily on different features in shaping human interactions and relations.

31 For two users i and j this framework proposes a generative probabilistic model for the observed data depending on the distance $\| \boldsymbol{\phi}^i - \boldsymbol{\phi}^j \|^2$, setting the probability of observing i interacting with j, denoted as $P(i \to j)$, as $P(i \to j)$ = logistic $\alpha - \beta \| \boldsymbol{\phi}^i - \boldsymbol{\phi}^j \|^2$ where α and β are shape parameters chosen to be 0.6 and 0.8 for the purposes of this illustration. These values are chosen to represent a sparse setting in which most individuals do not interact with one another.

32 The resulting interaction network drawn from the prescribed probability distribution does not exhibit the traditional properties of empirical social networks, which generally display the property of being sparse, clustered, with hubs, and scale free.

systems, shown in **Figure 4c**. These interactional traces constitute the ground truth of our recommendation problem.

In a recommender systems setting, the goal is to use observable data traces (*i.e.*, past choices made by users) to compute recommendations for new interactions. This setting represents social recommendations (*i.e.*, recommending users with which to interact), but the same formalization and operationalization applies for content recommendation. A recommender system is typically implemented to predict first the probability of observing an interaction, to then recommend those with high probability and that the user has not yet chosen. This is achieved by computing the probabilities of interactions on every potential pairwise interaction[33] to then rank them by probability to finally recommend only those ranked highest.

A traditional approach in recent AI recommender systems is to embed users and to rank recommen- dations on the bases of distances. The state of the art knows a wide multitude of embedding procedures for different types of data traces. We choose for this example an eigenvector approach (Greenacre 2017) for two reasons.

> First, this is a procedure successfully used in social media settings. Second, being a linear embedding procedure, it will allow us to highlight an important challenge of AI explainability: the complexity of the geometrical operations on the representation space on which recommendations depend, and that will be discussed in the next section.

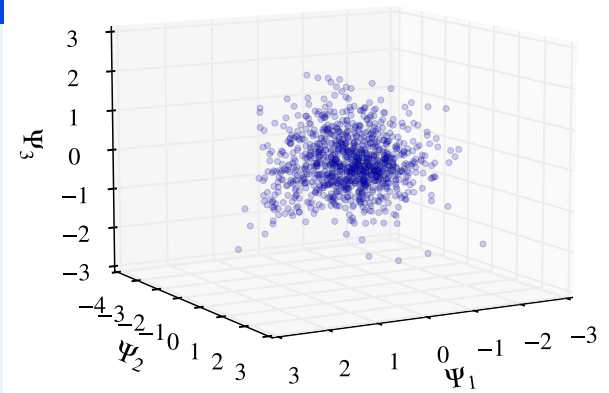
Concretely, we employ the embedding procedure described by Halford (2016) to embed the individuals of our population in a machine representation space Ψ of 64 dimensions.

33 Most modern recommender systems integrate procedures for limiting the scope of pairwise interactions (user-user or user-content) to consider, as N^2 potential (directed) interactions would result in $O(N^2)$ complexity, limiting industrial applications.

FIGURE 5.

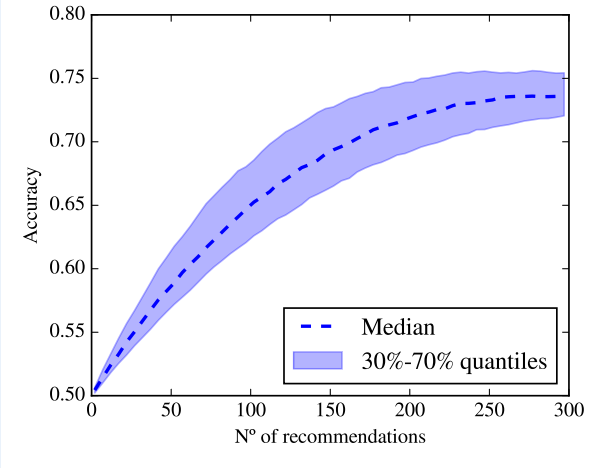
Representation learning space computed by an AI system along the the first three dimensions, and the accuracy of recommendations computed using this space for different numbers of propositions recommended to users.

5a.



5a. Positions of users along the first three dimensions of the representation learning space.

5b.



5b. Accuracy of recommendations made to users in the population by number of recommendations.

Figure 5a shows the positions of users in the synthetic along the first three of the 64 dimensions of the representation computed during training: Ψ_1, Ψ_2 and Ψ_3.

Using this machine spatial representation, recommendations are then computed by ranking pairwise distances as proxies for similarity in this new space. In this example, we intentionally avoid distinguishing between training and test set because we are not interested in arguing that this procedure produces accurate recommendations, let alone arguing that accuracy is competitive with regards to the state of the art. Instead, we take the accuracy of such a procedure as a baseline for comparing design strategies derived from AI explainability results to be proposed. We evaluate the accuracy of these recommendations using the *balanced accuracy metric* (Mower 2005):

$$\text{Accuracy} = \frac{\text{TPR} + \text{TNR}}{2}$$, where TPR stands for *true positive rate* and TNR stands for *true negative rate*. Concretely, for a number of recommendations made to a user, we count how many of those recommendations are *true positives* (*i.e.*, are also past choices observed in the training data), and how many of those recommendations are *true negatives* (i.e., how many of the interactions that were not recommended are also choices that the user did not make). We then compute TPR as TPR = TP/P where TP are true positives and P is the number of choices or interactions made by the users, and TNR as TNR = TN/N, where TN are *true negatives* and N is the number of choices or interactions that the user did not make.

Figure 5b shows the accuracy curve (for all users, showing median and 0.3 and 0.7 quantiles) for such a recommender system, in which recommending more options increases the accuracy. This accuracy profile will be taken as a baseline for comparison with accuracies produced by systems in which design strategies arising from explainability are applied.

5.2 MAPPING POLITICS IN REPRESENTATION LEARNING SPACES

In order to improve algorithm design with regards to their political impact through AI explainability, one needs to propose political hidden semantics for the representation learning space $\mathbf{\Psi}$. This involves leveraging knowledge about political opinions of individuals to provide meaning to machine spatial representations. While challenging in practice in real settings, in our synthetic data setting it is possible because the relevant features in our population were prescribed from the start. Let us consider the situation in which positions along the first dimension, $\mathbf{\Phi}_1$, stands for political positions. This does not change how the recommender system works, because it treats all dimensions alike. The question is whether we can use knowledge about positions along $\mathbf{\Phi}_1$ to map politics in the machine representation space $\mathbf{\Psi}$. Because of the embedding procedure chosen for recommendation, the expectation is that a multivariate linear regression should provide a good model for positions along $\mathbf{\Phi}_1$ as dependent variable using the 64 dimensions of $\mathbf{\Psi}$ as independent variables. If a multivariate linear regression model provides indeed a good model for $\mathbf{\Phi}_1$, we can also identify a spatial direction $\hat{\mathbf{\Phi}}_1$ in the representation learning space $\mathbf{\Psi}$ as the gradient of the linear model for $\mathbf{\Phi}_1$.

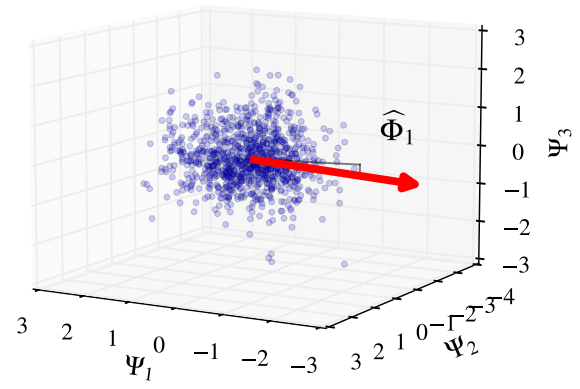
Figure 6a shows the three first dimensions of representation learning space $\mathbf{\Psi}$ with the direction $\hat{\mathbf{\Phi}}_1$ computed as a the gradient of the multivariate linear regression for $\mathbf{\Phi}_1$.

Figure 6b shows the quality of this multivariate regression model by plotting the known prescribed positions along $\mathbf{\Phi}_1$, and their values estimated via the regression, with mean l_1 error of 0.08 (for comparison, the prescribed variance along $\mathbf{\Phi}_1$ for the population is 1). This synthetic setting illustrates a concrete form of computation of hidden semantics for a chosen dimension of relevance in a population, illustrating a form of political AI explainability.

FIGURE 6.

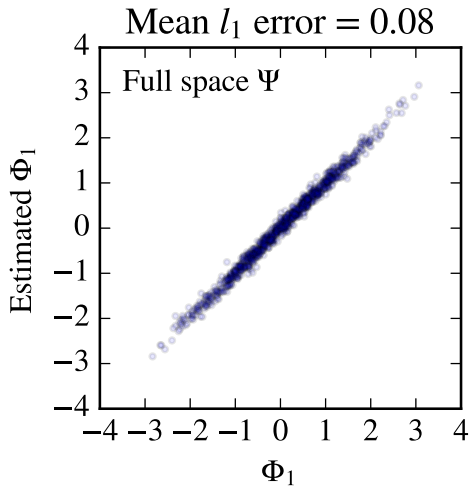
Estimation of spatial direction $\hat{\mathbf{\Phi}}_1$ (associated to political positions in the population $\mathbf{\Phi}_1$) in representation learning space $\mathbf{\Psi}$.

6a.



6a. Positions of users in representation learning space $\mathbf{\Psi}$, and estimated direction $\hat{\mathbf{\Phi}}_1$ related to political positions encoded in dimensions $\mathbf{\Phi}_1$ in the synthetic population.

6b.

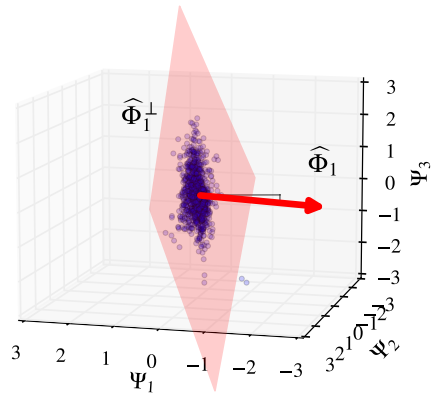


6b. Estimation of values along $\mathbf{\Phi}_1$ using a multivariate linear regression model with positions on representation learning space as independent variables.

FIGURE 7.

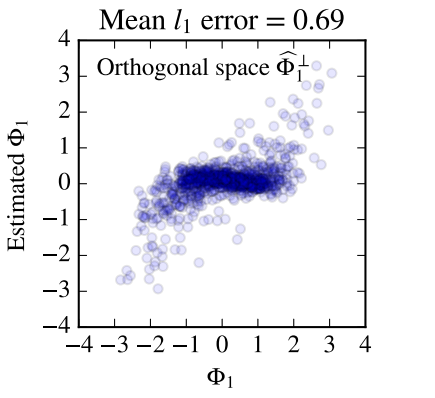
Representation learning space Ψ with identified political dimension $\hat{\Phi}_1$ allowing to restrict learning to the orthogonal space, rending recommendation blind to political information.

7a.



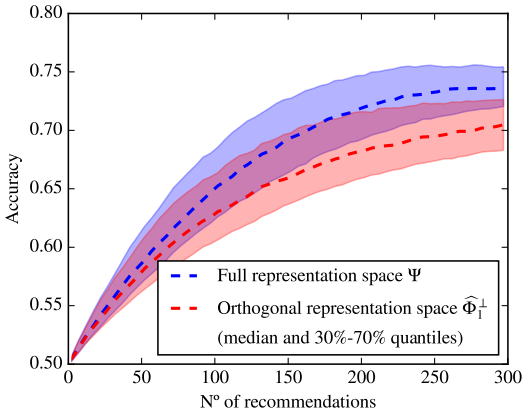
7a. Representation learning space Ψ, identified political dimension $\hat{\Phi}_1$, and orthongal subspace $\hat{\Phi}_1^{\perp}$ with population positions restricted to subspace.

7b.



7b. Loss of quality in multivariate regression models for political information $\hat{\Phi}_1$, showing the loss of political information in restricted learning in orthogonal subspace.

7c.



7c. Comparative accuracy between recommendations using the full representation learning space Ψ and the restricted orthogonal subspace $\hat{\Phi}_1^{\perp}$ that excludes political information. Selectively limiting information in representation learning may marginally diminishes accuracy.

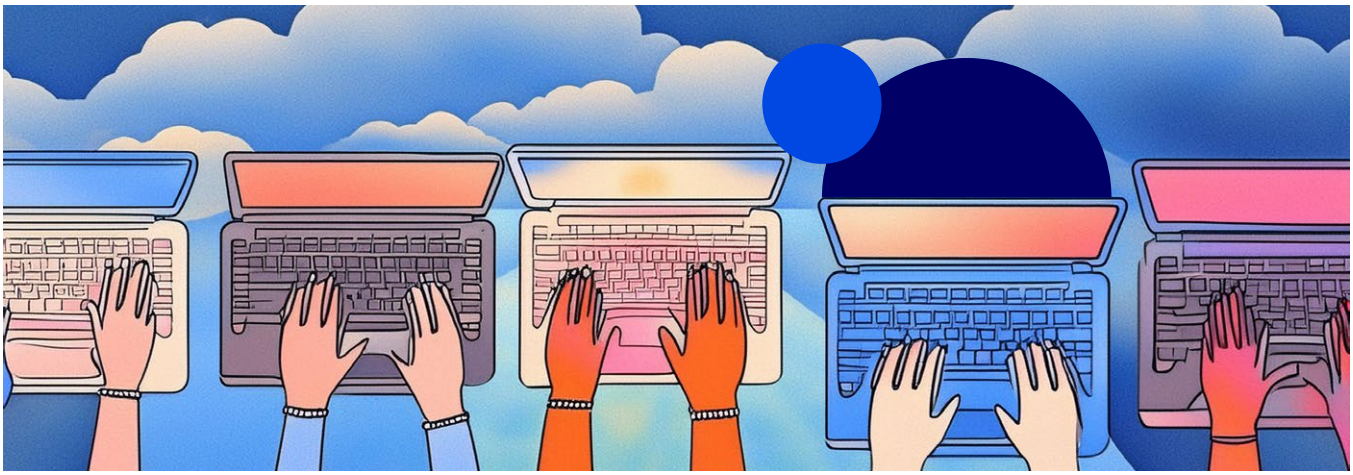
5.3 CONSTRAINING POLITICS IN AI LEARNING

The previous exercise constitutes AI explainability in the sense of hidden semantics described in the previous section, providing human intelligibility to the recommendation procedure involved in the results from **Figure 5**. Because this procedure is based on selection of nearest neighbors, the role of spatial direction $\hat{\Phi}_1$ in representation learning space Ψ, can be directly assessed. It is possible, for instance, to transform representation learning space to restrict how recommendations use information encoded in $\hat{\Phi}_1$. Let us suppose again that this dimension contains political information that we do not want the recommender system to use. If we identify $\hat{\Phi}_1$ in space Ψ, it is possible to consider its orthogonal subspace $\hat{\Phi}_1^{\perp}$ in representation space Ψ and the projections of all entities susceptible of being recommended (in this example users), thus limiting the computation of nearest neighbors to subspace $\hat{\Phi}_1^{\perp}$.

Figure 7a shows the direction $\hat{\Phi}_1$ in representation learning space Ψ and its orthogonal subspace $\hat{\Phi}_1^{\perp}$, with all entities now projected onto it.

Figure 7b shows the values of entities along Φ_1 (dependent variable) estimated using positions along $\hat{\Phi}_1^{\perp}$ (independent variables), the orthogonal space voided of information encoded in Φ_1. As expected, the quality or the regression is mostly lost, with mean l_1 error standing at 0.69 (compared to 0.08 using the full space Ψ). If we compute recommendations as nearest neighbors now on the orthogonal space $\hat{\Phi}_1^{\perp}$ (which contains less information than Ψ) the accuracy of recommendations is marginally diminished.

Figure 7c compares the accuracy curve for this new situation to the accuracy of recommendations using the full representation learning space.



In summary, this example used synthetic data to illustrate how features of users and contents online may be accessible to AI systems and encoded in representation learning space.

The previous sections discussed recent literature that showed this is possible for political opinions in real-world scenarios (without relying on synthetic data). This example also showed it is sometimes possible to map features of users and contents (or other entities operated in recommendations) in what is known as hidden semantics in AI explainability, enabling political explainability of AI systems whenever these features are linked to political positions. Most importantly, this example illustrates the operational details of the possibility of using this type of explainability in new design of algorithms, as presented in the work of Faverjon and Ramaciotti (2023). In this new paradigm of algorithmic design, when controlling for the effects of a feature such as political opinions, an alternative is proposed to the problem of AI risks regarding political segregation via selective exposure, and the moderation of potential outcomes such as polarization. Instead of taking a traditional normative approach specifying a diversity of recommendations (akin to taking nearest neighbors recommendations and modify them ex post to add diversity that might be lacking), hidden semantics explainable AI allows to cast the problem differently.

> **Instead of optimizing for accuracy and then introducing diversity of recommendations by prescribing how much diversity, the design principle here proposed aims at rendering the AI system agnostic to a particular feature.**

In this new formulation, the recommender is rendered blind to the feature of importance, at the cost of a loss in accuracy.

6. CONCLUSIONS: TOWARDS TOOLKITS AND GUIDELINES FOR AI POLICY AND REGULATION

This article discussed emerging opportunities brought about by the recent state of the art in AI systems for improving algorithmic moderation of digital public sphere. Because the digital public sphere plays at least a non-negligible role in the functioning of democracy (and arguably a consequent one in several countries), providers of AI services must monitor and improve these systems with regards to their risks. A prominent risk identified in the scientific literature and recent policy efforts (*e.g.*, the Digital Services Act) is that of systemic risks linked to selective exposure and political plurality. This is conceptualized in the scientific literature as political segregation, lack of diversity, and potentially connected with the risk of exacerbating political polarization.

The traditional approach to managing AI risks linked to political segregation in the digital public sphere (e.g., social media platforms) is to prescribe levels of recommended diversity, or to simultaneously optimize for accuracy and diversity during training (thus prescribing their relative importance). These approaches, however, incur the problems of any normative one:

- How much diversity should be prescribed?
- Who decides the diversity levels that are desirable or appropriate?
- How to decide the issue or political dimensions or categories along which diversity should be enforced?

This article also discussed additional risks of normative approaches to diversity; namely, the possibility that a sufficiently high level of political diversity in recommendations may also lead to exacerbated polarization. The article discusses recent experimental results that provide support to this consideration.

The opportunity for addressing this problem as discussed in this article is presented in more detail in the work of Faverjon and Ramaciotti (2023).

This article provides the substantial theoretical framework to link this proposition with

1. political science research addressing the question of how to conceptualize and operationalize political lines of divide of relevance in different social systems, as well as deciding which ones should be analyzed and addressed in algorithm design, and
2. policy and practitioner communities working on the impacts of AI in politics and democracy.

This proposed path to algorithm design draws from two theoretical perspectives:

1. recent representation learning AI systems that rely on spatial representation of data, and
2. spatial models of politics in social systems. Linking both on a theoretical and operational level, this article discussed the possibility of using the latter to produce AI explainability of the former (via a type of AI explainability techniques known as hidden semantics).

This article presented recent results showing

1. how it is possible to determine spatial political models of online social systems (through an empirical study focused on the French Twitter sphere),
2. how it is possible for algorithms to encode political information in representation learning procedures (through a study using the same empirical data and ubiquitous recommendation algorithms), and
3. how it is possible to use hidden semantics AI explainability hinging on political models to propose novel paths for the design of AI.

These novel paths propose ways to address these risks by selectively suppressing information that a recommender system may have learned, including, specially political positions of users and contents. Such a design principle would allow to address selective exposure based on politics, but without incurring normative approaches to political diversity. The concrete expressions of these design principles were illustrated with an example based on synthetic data, showing how to render an AI system blind to specific feature dimensions and the resulting impact in the achieved accuracy. It is important to remark that the result of such as design principle would remove the need for a normative prescription of diversity, but would affect

nonetheless the diversity perceived by users. A novel question would then be, whether this perceived diversity is greater or lesser than what would otherwise be recommended. While this important question has no simple answer, it must be remembered that the design principle does not to aim to set any given diversity level, but, on the contrary, to subtract the effect that the recommended diversity (either greater or lower) would have on users otherwise. This goal takes particular interest in light of recent research pointing to the danger of normative diversity in fostering exacerbated polarization.

> **To avoid the risks and downsides of normative approaches to political diversity in recommendations in social media, this article suggest, AI explainability hinging on political information that a machine may have learned must provide a way of rendering AI agnostic to the particular political dimensions that might be relevant for different national settings.**

Finally, the results discussed in this article point to an additional risk and ways to moderate it. Recent regulations limit the capacity of some digital services to process political information. Article 9 of the General Data Protection Regulation (GDPR) of the EU forbids the processing of sensitive data, which includes political opinions. Similarly, Article 26 of the Digital Services Act (DSA) of the EU forbids large digital service providers from recommending content based on profiles that contain sensitive information defined by Article 9 of the GDPR (thus including political opinions).

The results discussed in this article show that the risk exists that AI systems, specially representation learning ones, might use data traces to create profiles on sensitive data inadvertently and without this being an objective prescribed by designers. Concretely, it is possible for algorithms to treat data traces and create complex models for recommendation and that include information or profiles of users that contain quantifiable information on political opinions of data subjects, potentially in violation of Article 9 of the GDPR, or Article 25 of the DSA in some recommendation application. This identifies a new, previously unconsidered and credible compliance risk for digital services providers. The results and the opportunities here discussed also highlight a path for providers to self-assess compliance and avoid these risks.

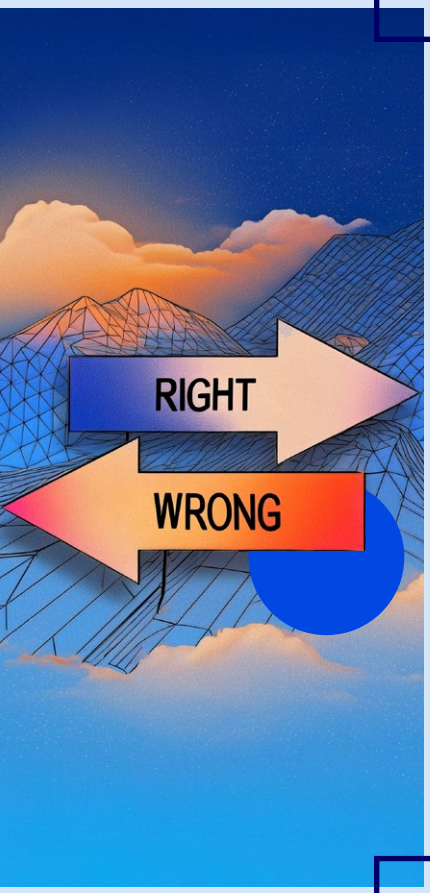
6.1 GUIDELINES

This discussion also highlighted several challenges in materializing the opportunities presented by the state of the art in AI, and that constitute the substance of the next subsection. Regardless, these results point to a number of potential actions that AI service providers may take to seize some of these opportunities.

1. **Production of markers enabling hidden semantics,** specially for political opinions, but also for other features relevant for moderation and compliance (e.g., those considered by GDPR, such as religious beliefs, ethnic origin, trade union membership and sexual orientation). Analyzing such information also comes with its own challenge regarding compliance with GDPR, to be discussed in the next subsection. However, as shown in the discussion of the results by Ramaciotti Morales et al. (2022), it is possible to compute hidden semantic hinging on political dimensions using markers that are not limited by GDPR. Concretely, political positions of political parties (but also that of news media outlets; Bakshy et al. 2015) may be available and free from limitations imposed by GDPR and other regulations. It is possible to use publicly-available datasets of positioning of entities along political dimensions to

1) assess the dimensions of relevance for a given national setting, and
2) to inspect representation learning spaces, providing quantitative measurements of political information.

Consequently, there is an advantage in specifying a list of markers (political, religious, ethnic, or other) by design, and that would enable algorithmic assessment via hidden semantic methods. In summary, providers of AI services should systematically identify markers (*e.g.*, political parties or personalities) with which to analyze embeddings spaces looking for inadvertently learned information.



2. Self-assessment of information learned by AI systems.
The need for self-assessment and reporting on AI risks is demanded by recent regulation and is set to become unavoidable task for sustainable digital service providers. This is stipulated for instance in Article 15 of the DSA, demanding providers to render publicly available yearly reports on moderation leading to compliance with the obligation set by the DSA. The results presented in this article suggest a path to improve compliance with these obligations by systematically self-assessing and measuring different information learned by AI systems that a service provider might have integrated. In doing so, there are approaches already suggested in the scientific literature, such as those from information theory. A main challenge to be discussed in this regard pertains to the geometrical complexity with which information might be encoded in representation learning.

In summary, AI security testing phases should benefit from systematic assessment of information learned using markers suited for different applications (*e.g.*, testing the presence of learned political information leveraging positions of political markers).

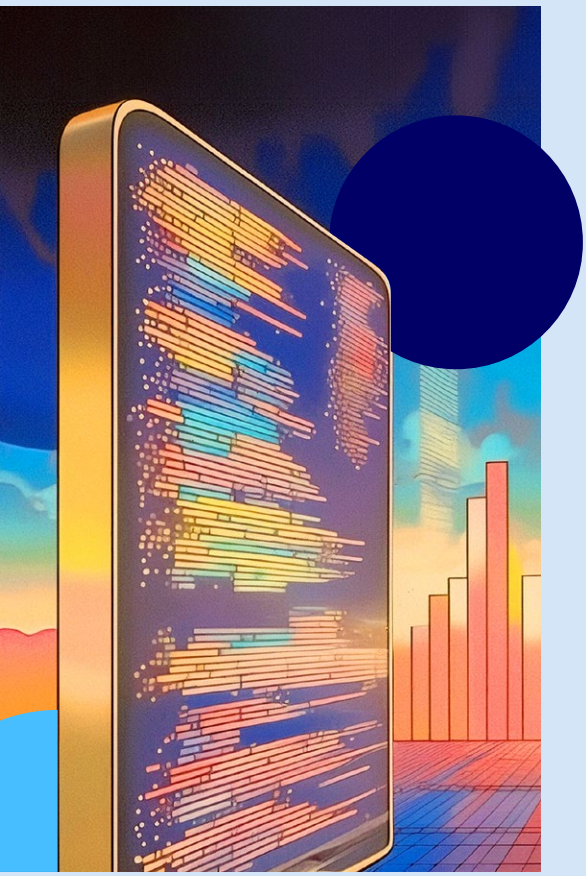
3. Evaluation of alternative representation learning procedures.
Similarly, through self- assessment, it is possible to continuously evaluate modifications to AI systems in line with the action presented in the data-driven example using synthetic data. Such continuous exploration has the potential of yielding better services both in terms of compliance and in terms of the impact of AI systems in society. The emergence of architectures that separate embedding from downstream machine tasks (*e.g.*, classification, regression), such as transformers, facilitates this exploration. An interesting alternative is to consider in the design phase safeguards assessing information in embeddings and limiting sensitive information when possible.

In summary, AI providers should systematically explore and assess (in terms of accuracy or other business-relevant metrics) modifications to the learned representation space, seeking to delete unwanted content (***e.g.***, political information), increasing the chances of compliance with regulation.

4. Design representation learning AI for openness.

Because of the identified risks of AI systems and their role in society, the scrutiny put on these systems and their providers is set to increase. This is in part reflected in recent regulation. Article 40 of the DSA, for instance, demands that large digital service providers grant access to data on the functioning of these services to regulators and researchers. The results of this article point to the interest of parties (platforms, service providers, regulators and researchers) in increasing collaboration in scrutiny of representation learning models. Such a collaboration demands that representation learning spatial models be designed in a way that enables the possibility of sharing them, potentially without revealing downstream operations that constitute the business opportunity of service providers.

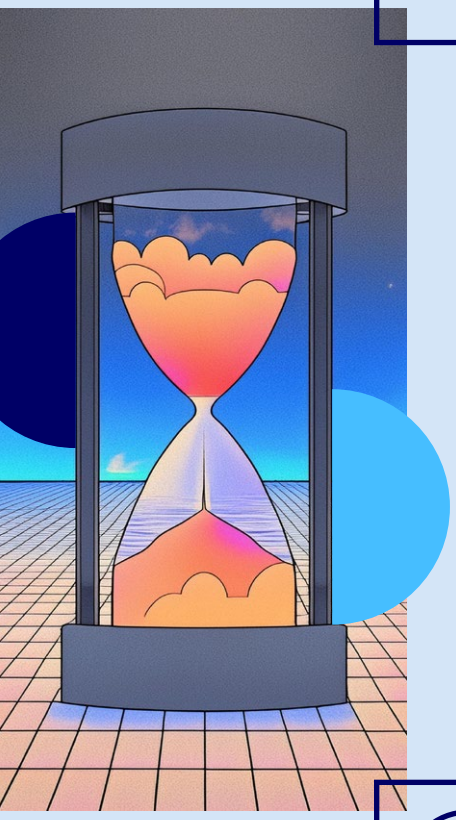
Additionally, the provisions of Article 40 of the DSA, in view of scientific research in AI explainability, may lead to openness obligations regarding trained models in the future, pointing to ways in which these providers may prepare. While DSA provides the best known example of obligations, the implementation of newer regulations such as the AI Act provide further examples motivating designing for openness.



In summary, the production of representation learning spaces, and that include markers for assessment, should be systematically included in the production of AI services as a means to facilitate sharing with external vetted researchers.

6.2 CHALLENGES

Finally, these results point to important challenges that are to be addressed in seizing the opportunities here outlined.



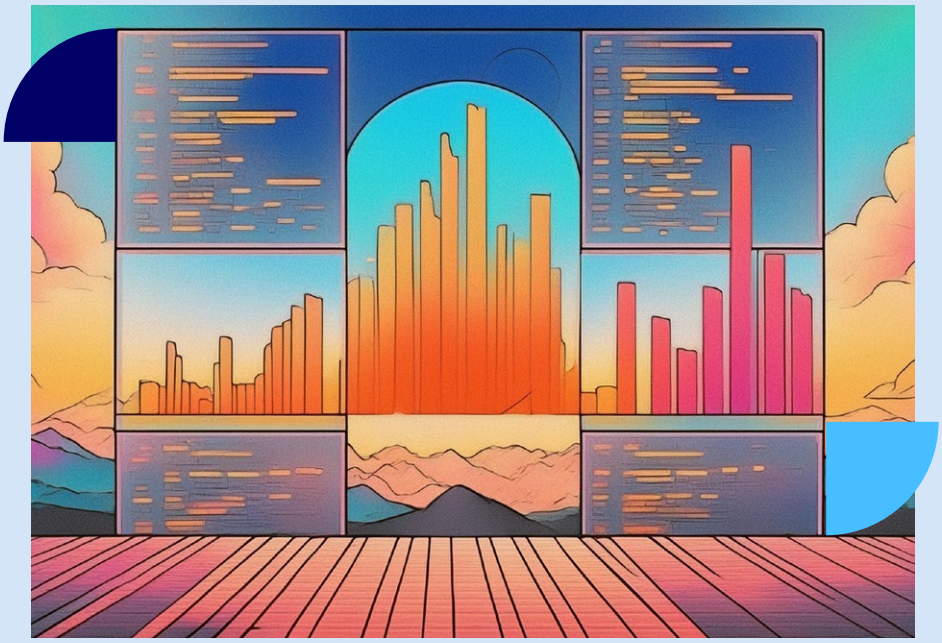
1. Geometrical complexity of learned representations.
The main scientific challenge in the program here outlined relates to the difficulty of proposing geometrical structures explaining the representation learning spaces and on which to predicate hidden semantics. This is best illustrated by the design choices taken in the previous section presenting the example of a synthetic population. While the recommendation principles adopted in that example are not foreign to industry, a growing part of AI systems used in digital services adopt deep learning methods that can accommodate a geometrically-complex encoding of information in weight spaces or embedding spaces, where the architectures allows it (*e.g.*, transformers). The family of systems presented in the previous section, in contrast, falls within those that comply with the linear representation hypothesis (Mikolov et al. 2013): systems in which properties of entities operated in machine representation (*e.g.*, users, contents) can be quantified in continuous scales which can be identified with spatial directions in representation learning spaces. This, of course, is not always the case for all AI systems, imposing additional complexity in describing non-linear encoding of properties in spaces, driving increasing research efforts.

2. Processing of sensitive (political opinion) data.
As mentioned in the previous section, producing analyses on AI explainability based on hidden semantics requires data descriptors on the populations or contents treated. Whenever these constitute sensitive data, the danger exist of incuring a violation of data protection regulation. This is a potential risk in advancing efforts in leveraging explainability in addressing risks linked to lack of diversity and segregation. From a compliance perspective (*i.e.*, assessing whether AI systems are treating sensitive information), this introduces an apparent paradox: *e.g.*, to assess whether an AI system is creating political profiles of individuals, the political profiles of individuals may be needed in the first place to evaluate machine representations. At least two solution may be given to this apparent paradox. One solution is to rely on openness by design and collaboration with actors free from limitations in data treatment, such as vetted academic researchers (a category

included, for example, in the DSA). A second solution is to rely on markers that are not subjected to limitations in profiling to examine representation learning spaces. This points again to the example of the use of political parties (but possibly also political figures and media outlets of known political tendencies) to inspect machine representations without having to rely on profiles of regular users.

3. Feature alignment and loss of information.

In a situation in which a political hidden semantic explanation has been developed for an AI system, it is possible to constrain the representation learning space to exclude political information in downstream tasks such as recommendation. In excluding this information, it is possible and even probable that other information will be also lost. Consider and idealized example in which political positions of users on a Left-Right scale is a linear function of age. Whenever political information is suppressed, information about age will also be rendered unavailable for recommendations. This has the potential of reducing accuracy, which is crucial to the business model sustaining digital services. Is it desirable or permissible to destroy profiling on non sensitive categories while seeking to destroy information on a particular sensitive category? The answer depends on the objective sought. If the imposition stems from the legal obligation to avoid profiling on a category, its suppression should be sought regardless of other information that may also be removed from the representation learning space.



REFERENCES

115th Congress of the US. 2017. "R h.r.4396—me too congress act". https://www.congress.gov/bill/ 115th-congress/house-bill/4396/text. Accessed: 2024-05-29.

Adadi, Amina and Mohammed Berrada. 2018. "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)". *IEEE access* 6 : 52138–52160.

Adamic, Lada A and Natalie Glance. 2005. "The political blogosphere and the 2004 us election: divided they blog". In *Proceedings of the 3rd international workshop on Link discovery*, pp. 36–43.

AI Act. 2021. "Regulation (eu) 2021/0106 of the european parliament and of the council laying down harmonised rules on artificial intelligence and amending certain union legislative acts". https://eur-lex.europa. eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206. Accessed: 2024-05-13.

Aiello, Luca Maria and Nicola Barbieri. 2017. "Evolution of ego-networks in social media with link recommendations". In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 111–120.

Ajzen, Icek. 1985. "From intentions to actions: A theory of planned behavior". *In Action contro*l, pp. 11–39. Springer.

Ajzen, Icek. 1989. "Attitude structure and behavior". *Attitude structure and function*.

Bail, Christopher A, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to opposing views on social media can increase political polarization". *Proceedings of the National Academy of Sciences 115 (37)*: 9216–9221.

Bail, Christopher A, Brian Guay, Emily Maloney, Aidan Combs, D Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky. 2020. "Assessing the russian internet research agency's impact on the political attitudes and behaviors of american twitter users in late 2017". *Proceedings of the national academy of sciences 117 (1)*: 243–250.

Bakshy, Eytan, Solomon Messing, and Lada A Adamic. 2015. "Exposure to ideologically diverse news and opinion on facebook". *Science* 348 (6239): 1130–1132.

Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. "Multimodal machine learning: A survey and taxonomy". *IEEE transactions on pattern analysis and machine intelligence* 41 (2): 423–443.

Barberá, Pablo. 2015. "Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data". *Political analysis*.

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost, and Joshua A Tucker. 2019. "Who leads? who follows? measuring issue attention and agenda setting by legislators and the mass public using social media data". *American Political Science Review* 113 (4): 883–901.

Barberá, Pablo, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. "Tweeting from left to right: Is online political communication more than an echo chamber?". *Psychological science*.

Bartolini, Stefano and Peter Mair. 2007. *Identity, competition and electoral availability: the stabilisation of European electorates 1885-1985*. ECPR Press.

Barzilai-Nahon, Karine. 2009. "Gatekeeping: A critical review". *Annual review of information science and technology* 43 (1): 1–79.

Bem, Daryl J. 1970. "Beliefs, attitudes, and human affairs."

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. 2013. "Representation learning: A review and new perspectives". *IEEE transactions on pattern analysis and machine intelligence* 35 (8): 1798–1828.

Benkler, Yochai, Robert Faris, and Hal Roberts. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bennett, W Lance and Barbara Pfetsch. 2018. *"Rethinking political communication in a time of disrupted public spheres"*. Journal of communication 68 (2): 243–253.

Benoit, Kenneth and Michael Laver. 2012. "The dimensionality of political space: Epistemological and methodological considerations". *European Union Politics* 13 (2): 194–218.

Bishop, Chris M. 1994. "Neural networks and their applications". *Review of scientific instruments* 65 (6): 1803– 1832.

Bobadilla, Jesus, Antonio Hernando, Fernando Ortega, and Jesus Bernal. 2011. "A framework for collaborative filtering recommender systems". *Expert Systems with Applications* 38 (12): 14609–14623.

Bobadilla, Jesús, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. "Recommender systems survey". *Knowledge-based systems* 46 : 109–132.

Bommasani, Rishi. 2023. "Ai spring? four takeaways from major releases in foundation models". *Stanford Institute for Human-Centered Artificial Intelligence. Retrieved from* https://hai. stanford. edu/news/ai-spring- four-takeaways-major-releases-foundation-models 30 : 2023.

Bond, Robert and Solomon Messing. 2015. "Quantifying social media's political space: Estimating ideology from publicly revealed preferences on facebook". *American Political Science Review* 19 .

Bondarenko, Andrey, Ludmila Aleksejeva, Vilen Jumutc, and Arkady Borisov. 2017. "Classification tree extraction from trained artificial neural networks". *Procedia Computer Science* 104 : 556–563.

Bonica, Adam. 2014. "Mapping the ideological marketplace". *American Journal of Political Science 58* (2): 367–386.

Bouchaud, Paul. 2024. "On meta's political ad policy enforcement: An analysis of coordinated campaigns & pro-russian propaganda".

Boutsidis, Christos and Efstratios Gallopoulos. 2008. "Svd based initialization: A head start for nonnegative matrix factorization". *Pattern recognition 41* (4): 1350–1362.

Budak, Ceren and Duncan J Watts. 2015. "Dissecting the spirit of gezi: Influence vs. selection in the occupy gezi movement". *Sociological Science* 2 : 370–397.

Buocz, Thomas, Sebastian Pfotenhauer, and Iris Eisenberger. 2023. "Regulatory sandboxes in the ai act: reconciling innovation and safety?". *Law, Innovation and Technology* 15 (2): 357–389.

Burkart, Nadia and Marco F Huber. 2021. "A survey on the explainability of supervised machine learning". *Journal of Artificial Intelligence Research* 70 : 245–317.

Cardon, Dominique, Jean-Philippe Cointet, and Antoine Mazières. 2018. "La revanche des neurones". *Réseaux* 211 (5): 173–220.

Castelvecchi, Davide. 2016. "Can we open the black box of ai?". *Nature News* 538 (7623): 20.

Chavalarias, David, Paul Bouchaud, and Maziyar Panahi. 2024. "Can a single line of code change society? the systemic risks for global information flow, opinion dynamics and social structures of recommender systems optimizing engagement". *Journal of Artificial Societies and Social Simulation*.

Chen, Qiwei, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. 2019. "Behavior sequence transformer for e-commerce recommendation in alibaba". In *Proceedings of the 1st international workshop on deep learning practice for high-dimensional sparse data*, pp. 1–4.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The statistical analysis of roll call data". *American Political Science Review* 98 (2): 355–370.

Cointet, Jean-Philippe, Pedro Ramaciotti Morales, Dominique Cardon, Caterina Froio, Andreı Mogoutov, Benjamin Ooghe-Tabanou, and Guillaume Plique. 2021. "What colours are the yellow vests? an ideological scaling of facebook groups". *Statistique et Société*.

Converse, Philip E. 1964. "The nature of belief systems in mass publics". *Ideology and Discontent*.

Dahl, Robert A. 1971. *Polyarchy*. Yale university press.

Dahl, Robert A. 2020. *On democracy*. Yale university press.

Dakhel, Arghavan Moradi, Vahid Majdinasab, Amin Nikanjam, Foutse Khomh, Michel C Desmarais, and Zhen Ming Jack Jiang. 2023. "Github copilot ai pair programmer: Asset or liability?". *Journal of Systems and Software* 203 : 111734.

Digital Services Act. 2022. "Regulation (eu) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services and amending directive 2000/31/ec". https://eur-lex.europa.eu/eli/reg/2022/2065/oj. Accessed: 2023-10-11.

Downs, Anthony. 1957. "An economic theory of political action in a democracy". *Journal of political economy* 65 (2): 135–150.

EUROSTAT. 2024. "59 percent of eu individuals using social networks in 2023". https://ec.europa.eu/ eurostat/web/products-eurostat-news/w/ddn-20240319-1, accessed on May 23, 2024.

Faverjon, Tim and Pedro Ramaciotti. 2023. "Discovering ideological structures in representation learning spaces in recommender systems on social media data". In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pp. 400–406.

Feinberg, Ayal, Regina Branton, and Valerie Martinez-Ebers. 2022. "The trump effect: how 2016 campaign rallies explain spikes in hate". *PS: Political Science & Politics* 55 (2): 257–265.

Gabielkov, Maksym, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. "Social clicks: What and who gets read on twitter?". In *Proceedings of the 2016 ACM SIGMETRICS international conference on measurement and modeling of computer science*, pp. 179–192.

Gallagher, Ryan J, Andrew J Reagan, Christopher M Danforth, and Peter Sheridan Dodds. 2018. "Di- vergent discourse between protests and counter-protests: #blacklivesmatter and #alllivesmatter". *PloS one* 13 (4): e0195644.

Gautam, Sanjana, Pranav Narayanan Venkit, and Sourojit Ghosh. 2024. "From melting pots to misrepresentations: Exploring harms in generative ai". *arXiv preprint arXiv:2403.10776*.

Gerber, Elisabeth R and Jeffrey B Lewis. 2004. "Beyond the median: Voter preferences, district heterogeneity, and political representation". *Journal of Political Economy* 112 (6): 1364–1383.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic content moderation: Technical and political challenges in the automation of platform governance". *Big Data & Society* 7 (1): 2053951719897945.

Greenacre, M. 2017. *Correspondence analysis in practice*.

Guess, Andrew M, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo- Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. 2023. "How do social media feed algorithms affect attitudes and behavior in an election campaign?". *Science 381* (6656): 398–404.

Habermas, Jürgen. 2015. *Between facts and norms: Contributions to a discourse theory of law and democracy*. John Wiley & Sons.

Halford, Max. 2016. "Prince". https://github.com/MaxHalford/prince.

Hix, Simon. 1999. "Dimensions and alignments in european union politics: Cognitive constraints and partisan responses". *European Journal of Political Research* 35 (1): 69–106.

Huszár, Ferenc, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. "Algorithmic amplification of politics on twitter". *Proceedings of the National Academy of Sciences* 119 (1): e2025334119.

Hyde, Susan D. 2020. "Democracy's backsliding in the international environment". *Science 369* (6508): 1192– 1196.

Imai, Kosuke, James Lo, Jonathan Olmsted, et al. 2016. "Fast estimation of ideal points with massive data". *American Political Science Review*.

Janssen, Davy and Raphaël Kies. 2005. "Online forums and deliberative democracy". *Acta política* 40 : 317–335.

Jolly, Seth, Ryan Bakker, Liesbet Hooghe, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. 2022. "Chapel hill expert survey trend file, 1999–2019". *Electoral Studies* 75 : 102420.

Jost, John T, Delia S Baldassarri, and James N Druckman. 2022. "Cognitive–motivational mechanisms of political polarization in social-communicative contexts". *Nature Reviews Psychology* 1 (10): 560–576.

Jungherr, Andreas. 2016. "Twitter use in election campaigns: A systematic literature review". *Journal of information technology & politics* 13 (1): 72–91.

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

Jurafsky, Daniel and James H Martin. 2022. *Speech and Language Processing* (3rd edition ed.). Prentice Hall.

Kalogeropoulos, Antonis, Jane Suiter, Linards Udris, and Mark Eisenegger. 2019. "News media trust and news consumption: Factors related to trust in news in 35 countries". *International journal of communication* 13 : 22.

Krenn, Mario, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, et al. 2022. "On scientific understanding with artificial intelligence". *Nature Reviews Physics* 4 (12): 761–769.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. "Imagenet classification with deep convolutional neural networks". *Advances in neural information processing systems* 25 .

Kulshrestha, Juhi, Muhammad Zafar, Lisette Noboa, Krishna Gummadi, and Saptarshi Ghosh. 2015. "Characterizing information diets of social media users". In *Proceedings of the international AAAI conference on web and social media*, Volume 9, pp. 218–227.

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. 2010. "What is twitter, a social network or a news media?". In *Proceedings of the 19th international conference on World wide web*, pp. 591–600.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting policy positions from political texts using words as data". *American political science review* 97 (2): 311–331.

Lazarsfeld, Paul F, Robert K Merton, et al. 1954. "Friendship as a social process: A substantive and methodological analysis". *Freedom and control in modern society* 18 (1): 18–66.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning". ***nature*** 521 (7553): 436–444.

Lewis, Jeffrey B and Gary King. 1999. "No evidence on directional vs. proximity voting". *Political analysis* 8 (1): 21–33.

Lim, Merlyna. 2012. "Clicks, cabs, and coffee houses: Social media and oppositional movements in egypt, 2004–2011". *Journal of communication* 62 (2): 231–248.

Lipset, Seymour Martin and Stein Rokkan (Eds.)1967. ***Party Systems and Voter Alignments: Cross-national Perspectives***, Volume 7. New York: Free Press.

Liu, Bing and Lei Zhang. 2012. "A survey of opinion mining and sentiment analysis". In *Mining text data*, pp. 415–463. Springer.

Liu, Ruibo, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. "Quantifying and alleviating political bias in language models". *Artificial Intelligence* 304 : 103654.

Lorenz-Spreen, Philipp, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. 2023. "A systematic review of worldwide causal and correlational evidence on digital media and democracy". *Nature human behaviour* 7 (1): 74–101.

Luo, Xin, Mengchu Zhou, Yunni Xia, and Qingsheng Zhu. 2014. "An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems". *IEEE Transactions on Industrial Informatics* 10 (2): 1273–1284.

MacAvaney, Sean, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. "Hate speech detection: Challenges and solutions". *PloS one* 14 (8): e0221152.

Marcinkevičs, R. and J. E. Vogt. 2020. "Interpretability and Explainability: A Machine Learning Zoo Mini-tour".

Marks, Gary, David Attewell, Liesbet Hooghe, Jan Rovny, and Marco Steenbergen. 2022. "The social bases of political parties: A new measure and survey". *British Journal of Political Science*: 1–12.

Marquez, Leorey, Tim Hill, Reginald Worthley, and William Remus. 1991. "Neural network models as an alternative to regression". In *Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences,* Volume 4, pp. 129–135. IEEE.

Martínez-Plumed, Fernando, Pablo Barredo, Sean O Heigeartaigh, and Jose Hernandez-Orallo. 2021. "Research community dynamics behind popular ai benchmarks". *Nature Machine Intelligence* 3 (7): 581–589.

McCoy, Jennifer and Murat Somer. 2019. "Toward a theory of pernicious polarization and how it harms democracies: Comparative evidence and possible remedies". *The Annals of the American Academy of Political and Social Science* 681 (1): 234–271.

McGaughey, Ewan. 2022. "Will robots automate your job away? full employment, basic income and economic democracy". *Industrial Law Journal* 51 (3): 511–559.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook. 2001. “Birds of a feather: Homophily in social networks”. *Annual Review of Sociology* 27 (1): 415–444.

Messaoudi, Chaima, Zahia Guessoum, and Lotfi Ben Romdhane. 2022. “Opinion mining in online social media: a survey”. *Social Network Analysis and Mining* 12 (1): 1–18.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. “Distributed representations of words and phrases and their compositionality”. *Advances in neural information processing systems* 26 .

Morozov, Evgeny. 2011. *The net delusion: How not to liberate the world*. Penguin UK.

Mower, Jeffrey P. 2005. “Prep-mt: predictive rna editor for plant mitochondrial genes”. BMC *bioinformatics* 6 : 1– 15.

Norvig, P and S Russel. 2002. “Artificcial intelligence: A modern approach”. Prentice Hall Upper Saddle River, NJ, USA: Rani, M., Nayak, R., & Vyas, OP (2015). *An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage*. Knowledge-Based Systems 90 : 33–48.

Olbrich, Eckehard and Sven Banisch. 2021. “The rise of populism and the reconfiguration of the german political space”. *Frontiers in big Data:* 83.

Pariser, Eli. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Parsons, Talcott, Robert Freed Bales, and Edward Shils. 1953. *Working papers in the theory of action*. The Free Press.

Peress, Michael. 2022. “Large-scale ideal point estimation”. *Political Analysis* 30 (3): 346–363.

Poole, Keith T and Howard Rosenthal. 1985. “A spatial model for legislative roll call analysis”. *American Journal of Political Science*.

Poole, Keith T and Howard Rosenthal. 2000. *Congress: A political-economic history of roll call voting*. Oxford University Press on Demand.

Przeworski, Adam. 2018. *Why bother with elections?* John Wiley & Sons.

Ramaciotti Morales, Pedro, Manon Berriche, and Jean-Philippe Cointet. 2023. “The geometry of misinformation: embedding twitter networks of users who spread fake news in geometrical opinion spaces”. In *International Conference on Web and Social Sciences* ICWSM. AAAI.

Ramaciotti Morales, Pedro and Jean-Philippe Cointet. 2021. “Auditing the effect of social network recommenda- tions on polarization in geometrical ideological spaces”. In *15th ACM Conference on Recommender Systems*, RecSys’ 21.

Ramaciotti Morales, Pedro, Jean-Philippe Cointet, Bilel Benbouzid, Dominique Cardon, Caterina Froio, Omer Faruk Metin, Benjamin Ooghe, and Guillaume Plique. 2021. “Atlas multi-plateformes d’un mouvement social: Le cas des gilets jaunes”. *Statistique et Société*.

Ramaciotti Morales, Pedro, Jean-Philippe Cointet, and Caterina Froio. 2022. “Posters and protesters: The networked interplay between onsite participation and facebook activity in the yellow vests movement in france”. *Journal of Computational Social Science* 5 (2): 1129–1157.

Ramaciotti Morales, Pedro, Jean-Philippe Cointet, and Gabriel Muñoz Zolotoochin. 2021. “Unfolding the dimensionality structure of social networks in ideological embeddings”. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 333–338.

Ramaciotti Morales, Pedro, Jean-Philippe Cointet, Gabriel Muñoz Zolotoochin, Antonio Fernández Peralta, Gerardo Iñiguez, and Armin Pournaki. 2022. “Inferring attitudinal spaces in social networks”. *Social Network Analysis and Mining* 13 (1): 14.

Ramaciotti Morales, Pedro and Zografoula Vagena. 2022. “Embedding social graphs from multiple national settings in common empirical opinion spaces”. In *Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

Riker, William H and Peter C Ordeshook. 1968. “A theory of the calculus of voting”. *American political science review* 62 (1): 25–42.

Roth, Camille, Antoine Mazières, and Telmo Menezes. 2020. “Tubes and bubbles topological confinement of youtube recommendations”. *PloS one* 15 (4): e0231703.

Rozado, David. 2023. “The political biases of chatgpt”. *Social Sciences* 12 (3): 148.

Satuluri, Venu, Yao Wu, Xun Zheng, Yilei Qian, Brian Wichers, Qieyun Dai, Gui Ming Tang, Jerry Jiang, and Jimmy Lin. 2020. “Simclusters: Community-based representations for heterogeneous recommendations at twitter”. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3183–3193.

Schafer, J Ben, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. “Collaborative filtering recommender systems”. In *The adaptive web: methods and strategies of web personalization*, pp. 291–324. Springer.

Schradie, Jen. 2011. “The digital production gap: The digital divide and web 2.0 collide”. *Poetics* 39 (2): 145–168.

Scott, A Carlisle, William J Clancey, Randall Davis, and Edward H Shortliffe. 1977. “Explanation capabilities of production-based consultation systems”. *American Journal of Computational Linguistics*: 1–50.

Shirky, Clay. 2009. *Here comes everybody: How change happens when people come together*. Penguin UK.
Shoemaker, Pamela J and Timothy Vos. 2009. Gatekeeping theory. Routledge.

Slapin, Jonathan B and Sven-Oliver Proksch. 2008. “A scaling model for estimating time-series party positions from texts”. *American Journal of Political Science* 52 (3): 705–722.

Sonnett, John. 2004. “Musical boundaries: intersections of form and content”. *Poetics* 32 (3-4): 247–264.

Sunstein, Cass R. 2001. *Republic.com*. Princeton university press.

Teorell, Jan, Michael Coppedge, Staffan Lindberg, and Svend-Erik Skaaning. 2019. “Measuring polyarchy across the globe, 1900–2017”. *Studies in Comparative International Development* 54 : 71–95.

Tufekci, Zeynep and Christopher Wilson. 2012. “Social media and the decision to participate in political protest: Observations from tahrir square”. *Journal of communication* 62 (2): 363–379.

Uscinski, Joseph E, Adam M Enders, Michelle I Seelig, Casey A Klofstad, John R Funchion, Caleb Everett, Stefan Wuchty, Kamal Premaratne, and Manohar N Murthi. 2021. “American politics in two dimensions: Partisan and ideological identities versus anti-establishment orientations”. *American Journal of Political Science 65* (4): 877–895.

V-Dem Institute. 2019. “Varieties of democracy (v-dem) annual report 2019-” democracy facing global challenges””.

Vafa, Keyon, Suresh Naidu, and David M Blei. 2020. “Text-based ideal points”. *arXiv preprint arXiv:2005.04232*.

Vasconcelos, Vítor V, Sara M Constantino, Astrid Dannenberg, Marcel Lumkowsky, Elke Weber, and Simon Levin. 2021. “Segregation and clustering of preferences erode socially beneficial coordination”. *Proceedings of the National Academy of Sciences* 118 (50): e2102153118.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention is all you need”. *Advances in neural information processing systems 30*.

Wu, Patrick Y, Joshua A Tucker, Jonathan Nagler, and Solomon Messing. 2023. “Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting”. *arXiv preprint arXiv:2303.12057*.

Zhang, Yu, Peter Tiňo, Aleš Leonardis, and Ke Tang. 2021. “A Survey on Neural Network Interpretability”. 5 (5): 726–742.

Zhou, Tao, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. “Solving the apparent diversity-accuracy dilemma of recommender systems”. *Proceedings of the National Academy of Sciences 107* (10): 4511–4515.

Ziegler, Cai-Nicolas, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. “Improving recommendation lists through topic diversification”. In *Proceedings of the 14th international conference on World Wide Web,* pp. 22–32.

IE UNIVERSITY
CENTER FOR THE GOVERNANCE OF CHANGE

WRITTEN BY
Pedro Ramaciotti

This paper is part of a series of four papers within AI4Democracy, a global research and outreach initiative led by the Center for the Governance of Change at IE University, with Microsoft as strategic supporter. AI4Democracy seeks to harness AI to defend and strengthen democracy through coalition-building, advocacy, and intellectual leadership.

SUGGESTED CITATION
Ramaciotti, P. (2024). *Depolarizing and moderating social media with AI: Tools and guidelines leveraging representation spaces*, AI4Democracy, IE Center for the Governance of Change.

ACKNOWLEDGEMENTS
The author acknowledges financial support from the AI4Democracy project from IE University, Spain.

Images: All images were generated by different AI tools.
Design: epoqstudio.com

FOR MORE INFORMATION ON THE AI4DEMOCRACY INITIATIVE, VISIT:

IE.EDU/CGC/RESEARCH/AI4DEMOCRACY