

RECOMMENDER SYSTEMS AND
DEMOCRACY: LINGUISTIC DIVERSITY,
EPISTEMIC AGENCY, AND CIVIC
PARTICIPATION

A thesis submitted to
Trinity College Dublin, the University of Dublin

for the degree of
Doctor in Philosophy

in the
School of Languages, Literatures and Cultural Studies

by
Artur Osipov

Supervisor: Professor Omar Garcia
Co-Supervisor: Professor Jennifer Edmond

2026

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I confirm that the work has been completed in full compliance with relevant university policies, including the Academic Integrity policy, the Trinity Policy on Good Research Practice, and the guidance on AI and GenAI in Teaching, Learning, Assessment and Research.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

A handwritten signature in black ink, appearing to read 'Artur Osipov', with a large, stylized initial 'A'.

Signed:

Artur Osipov

Summary

This thesis examines how recommender systems (RS) affect democracy by reshaping linguistic diversity and epistemic agency. It addresses the underexplored question of how RS influence language use and knowledge practices. The study analyses RS in platform environments, considers how they are framed by policymakers, and investigates how they are represented in research discourse. The main objectives were threefold. First, to examine how RS shape the language and perspectives that users of video-hosting platforms are exposed to. Second, to compare how policymakers conceptualise and address AI and how this affects RS. Third, to analyse how RS are framed in research discourse and what assumptions guide that framing. Together, these aims link empirical, policy, and research perspectives to clarify mechanisms through which RS affect democracy. The thesis employed qualitative and comparative methods across three strands. 1) Two case studies of YouTube and Dzen were conducted using sock puppet accounts with specific queries to capture recommendation outputs. These datasets were analysed through critical discourse analysis to trace changes in linguistic range and sentiment distribution. 2) A comparative analysis examined AI policy frameworks in the EU, the US, and Russia through official documents, policy papers, and strategies. 3) A third strand analysed research publications on RS, mapping problem framings, proposed solutions, and the extent to which social and linguistic concerns are addressed. The three strands were integrated through comparative design, connecting platform practices, policy approaches, and disciplinary discourses.

The analysis shows three main findings. First, the case studies indicate that RS narrow linguistic diversity by amplifying repeated patterns and dominant discourses, limiting opportunities for interpretation and reducing epistemic agency. Second, the policy comparison reveals contrasting approaches: the EU foregrounds risks and rights; the US prioritises markets and innovation; Russia centres control and information management. Each shapes RS oversight in different ways, pointing to the absence of a common democratic framework. Third, the research discourse is dominated by technical framings and solutionist perspectives that marginalise social and linguistic dimensions. These findings show that RS function within wider governance and research contexts that influence democratic possibilities. Building on these findings, the thesis positions RS as cultural technologies of omission. Chapter 6 situates RS alongside other technologies that operate through narrowing, excluding, or filtering, placing them within a broader historical and cultural lineage. This perspective highlights omission as a constitutive feature of RS as cultural-technical systems. The thesis contributes in four ways. First, it connects established concerns about RS and polarisation with new evidence of their linguistic effects in English and Russian. Second, it shows that RS operate through similar patterns of omission in both democratic and authoritarian contexts, despite different political and economic drivers. Third, it demonstrates how discourse about RS separates researchers, policymakers, and users, shaping what is understood and acted upon. Fourth, it argues that RS demand greater attention as cultural technologies with significant implications for democratic processes, highlighting design practices such as loss functions that are rarely examined through the lens of omission.

Acknowledgement

I am grateful to my supervisor, Professor Jennifer Edmond, for her support throughout this research and beyond. Over these four years I have faced changes in my academic practice and in navigating the challenges of becoming a refugee. Jennifer's guidance, empathy, and encouragement sustained me, and I will always remain thankful for her presence. I also thank my supervisor, Professor Omar García, for his formal support of this project. This research was made possible through the Provost's PhD Project Award, which enabled me to apply my perspective and background to the study of AI's impact on democracy.

I am deeply grateful to Raimonda Elvikyte for her professionalism and constant support, which made navigating this PhD both smooth and reassuring. She is a reminder that systems function because of the dedication and care of the people behind them. I also thank Professor Konrad Kording for shaping my engagement with loss functions, and my collaborators and colleagues Dr Doireann Wallace and Dr Dharm Kapletia, with whom it has been a joy to work.

My thanks go to my soulmate, Elizaveta Vorobeva, for her love, patience, and art, which sustained me throughout this project. I am also grateful to my sister, Dr Elena Osipova, whose own difficult PhD journey reminded me that mine could always be worse, and so became a source of motivation. I thank my friend Ilya Cekinov, whose encouragement brought me to Ireland and into this chapter of my life. Finally, I thank Wendy Lyon for her unwavering support and for being there through it all.

Table Of Contents

Summary	i
Acknowledgement	iii
Table Of Contents	iv
List of Figures	v
List of Tables	vii
Chapter 1. Introduction	1
Chapter 2. Literature review	33
Chapter 3. Recommender Systems: Case Study	98
3.1 Methodology	98
3.2 US Case Study	116
3.3 Russia Case Study	143
3.4 Conclusion	160
Chapter 4. Recommender Systems: AI Policy Frameworks	164
4.1 Methodology	165
4.2 Technology: To what extent do AI definitions in the EU, US, and Russia explicitly frame RS as a distinct technical object?	173
4.3 Economy: How do the economic priorities implicit in AI policies influence the regulation of RS?	192
4.4 Culture: Do current AI policies operationalise fairness, transparency, adaptability, and trust for RS?	201
4.5 Quality: Do current AI policies operationalise fairness, transparency, adaptability, and trust for RS?	212
4.6 Conclusion	217
Chapter 5. Recommender systems: Research Discourse	219
5.1 Methodology	219
5.2 Governance & Accountability	228
5.3 Agency & Autonomy	243
5.4 Technical Instrumentalism	253
5.5 Risk & Mitigation	262
5.6 Conclusion	276
Chapter 6: Epistemic Legacies in Algorithmic Research	281
Chapter 7: Conclusion	332
Bibliography	357

List of Figures

Figure 1 Sock puppet categories, associated queries, and platforms.	106
Figure 2 TTR and vocabulary size for USB and USR.....	117
Figure 3 TTR and vocabulary size by age groups in US case study.	118
Figure 4 Average positive and negative sentiment scores by gender and age group for USB and USR.	130
Figure 5 Compound sentiment scores for USB and USR.....	137
Figure 6 TTR and vocabulary size for RUB and RUR.	145
Figure 7 TTR and vocabulary size by age groups in Russia Case Study.....	146
Figure 8 Average positive and negative sentiment scores by gender and age group for RUB and RUR.	150
Figure 9 Compound sentiment scores for RUB and RUR.	157
Figure 10 Normalised per 10k words frequency comparison for the recommendation, bias, and transparency.....	186
Figure 11 Raw frequency of economic terms compliance, market, cost, and investment.	198
Figure 12 Top five most frequent terms in the Governance and Authority.	203
Figure 13 Top five most frequent terms in the Individual versus Collective Orientations category.....	205
Figure 14 Top five most frequent terms in the Risk and Security category. ...	206
Figure 15 Top five most frequent terms in the Values and Social Harmony category.	208
Figure 16 Normalised per-million growth of key governance-related terms in the English and Russian corpora.....	234

Figure 17 Normalised frequencies per 10k words of top co-occurring terms for transparency and regulation in the English corpus, and ‘прозрачность’ [transparency] and ‘регулирование’ [regulation] in the Russian corpus.	236
Figure 18 Normalised per-million growth of key Agency & Autonomy related terms in the English and Russian corpus.	248
Figure 19 Top co-occurring terms for agency, autonomous, and autonomy..	249
Figure 20 Normalised per-million growth of key Technical terms in the English and Russian corpus.	257
Figure 21 Top co-occurring terms for algorithm, model, technology, and computation.	258
Figure 22 Top co-occurring terms for ‘алгоритм’ [algorithm], ‘модель’ [model], ‘технология’ [technology], and ‘реализация’ [computation].	259
Figure 23 Normalised per-million growth of key Risk & Mitigation terms in the English corpus and in the Russian corpus.	267
Figure 24 Top co-occurring terms for risk, harm, safety.	269
Figure 25 Top co-occurring terms for likelihood, trust, and ethical.	269
Figure 26 Top co-occurring terms for ‘риск’ [risk], ‘вред’ [harm], ‘безопасность’ [safety].	270
Figure 27 Top co-occurring terms for ‘вероятность’ [likelihood], ‘доверие’ [trust], and ‘этический’ [ethical].	271
Figure 28 Reinforcing cycle of RS epistemic control.	331

List of Tables

Table 1 Keyword sets for the combined cultural framework categories.	171
Table 2 Raw frequencies of top co-occurring terms for transparency, accountability, governance, and regulation in English- and Russian-language corpora (2010–2024).	237

Chapter 1. Introduction

In 2019 Kevin Roose published an article in *The New York Times* entitled 'The Making of a YouTube Radical,' stating that 'YouTube has inadvertently created a dangerous on-ramp to extremism by combining two things: a business model that rewards provocative videos with exposure and advertising dollars, and an algorithm that guides users down personalized paths meant to keep them glued to their screens' (Roose, 2019). Roose interviews Mr. Cain, who 'recently swore off the alt-right nearly five years after discovering it, and has become a vocal critic of the movement,' and discusses his 'radicalisation' through YouTube recommender system (RS). The article shows how content selected by an artificial intelligence (AI)-driven technology affected the diversity of knowledge communicated to him and, in effect, 'brainwashed' him. Roose concludes that Mr. Cain 'had successfully climbed out of a right-wing YouTube rabbit hole, only to jump into a left-wing YouTube rabbit hole.' Even though 'YouTube's RS is not set in stone' and 'small changes' are made every year, this case emphasises how the ability of technology to influence perception and discourse warrants examination.

The case of Mr. Cain illustrates how AI-driven systems affect the knowledge available to citizens and shape the nature of their civic participation, as he not only absorbed far-right content but also adopted new political labels, argued with friends, produced his own videos recounting his radicalisation, and later launched a left-wing YouTube channel that positioned him as an active participant in public discourse (Roose, 2019). It raises questions about perceiving such effects when they do not push individuals to political extremes

and, as a result, remain less visible, since gradual shifts in attitudes, language, and everyday practices are harder to detect than overt radicalisation. This dissertation examines how contemporary AI systems that operate with text and language can structure conditions of civic participation and epistemic agency, using recommender systems as the primary example. The emphasis is on cultural rather than overtly political forms of participation, such as voting. Cultural participation on online platforms is associated with political participation and civic engagement as part of a complex interaction of personal, structural, and cultural factors (Alex-Assensoh, 2005: p. 52). While political and social participation has intrinsic value for human life and well-being (Sen, 2017), cultural participation can influence individual behaviour and shape identity (Neisser & Jopling, 1997), particularly as more people engage in content creation rather than mere consumption (Halpern & Gibbs, 2013). Social media may not provide a forum for in-depth policy debate, but it functions as a deliberative space that can encourage political participation, both directly and indirectly (p. 1166). In political contexts where complex issues require rational discourse, deep learning AI systems that curate content on platforms can influence patterns of exposure to information that may condition perceptions of public opinion and facts. This process can create echo chambers that reinforce prior convictions and intensify societal divides, as the algorithmic structuring of information privileges engagement-driven content over deliberative exchange and narrows the range of perspectives encountered (Cho et al., 2020). On YouTube, related video recommendations are 'the most important view source for the majority of videos' (Zhou, Khemmarat & Gao, 2010). This reinforces the Matthew effect in platform RS, where the system prioritises popular and frequently viewed videos (Liu &

Huang, 2021). Such ranking does not necessarily align with content relevance and highlights the need to consider the interests of multiple stakeholders embedded in RS environments, including platform providers, advertisers, users as participants, and content producers (Milano, Taddeo & Floridi, 2020).

In contemporary platform ecosystems, however, these roles are increasingly interwoven. Users are not merely consumers of recommendations but also active contributors to the informational environment through practices of commenting, sharing, rating, and content creation. Following Ritzer, Dean, and Jurgenson's concept of the *prosumer*, production and consumption are no longer analytically separable, particularly in Web-based media systems where users simultaneously generate data, create content, and consume algorithmically curated outputs (Ritzer, Dean & Jurgenson, 2012). Content creators operating within such systems are likewise shaped by recommender logics, as visibility, monetisation, and cultural relevance are conditioned by engagement metrics and algorithmic optimisation. As a result, creators frequently adapt linguistic style, thematic focus, and output frequency to align with recommender incentives, effectively co-producing the informational environment together with the platform and its users as participants. Given this structural entanglement, this dissertation treats users and content producers not as discrete stakeholder categories but as constituent actors within society whose participation unfolds within environments structured by recommender systems, where visibility, relevance, and circulation are shaped by algorithmic ranking rather than individual intent alone (Ritzer, 2015). This approach allows analysis to focus on how RS configure the conditions of participation and knowledge circulation across the social field,

rather than attributing agency or responsibility to isolated actor groups within an algorithmically mediated environment.

Engineered sociality: AI and society/information society

The concept of the 'information society' is used in this work to frame the discussion. Before examining the impact of AI on democracy, it is necessary to define the broader context in which social actors engage with these systems as a set of technologies, shaped by data, algorithms, and computational infrastructures, and the factors influencing their development. Social capital and its role in shaping a democratic environment make identifying the type of society under discussion relevant, since the characteristics of an information society differ from those of industrial or pre-digital societies (Frank, 2005: p. 2). The 'public good aspect' of social capital 'leads to underinvestment in' it (Coleman, 1988). Its capacity to support democracy by providing 'a space for the creation and dissemination of discourse critical of the present government' and 'a way for active opposition to the regime to grow' makes the relationship with democracy 'reciprocal' (Paxton, 2002). AI-driven technologies can shape conditions for cultural participation that are associated with forms of social capital, thereby constraining or enabling patterns of civic engagement.

The current state of society is highly dependent upon the implementation of contemporary information and communication technologies (ICTs), due to the large amount of information produced and processed daily by media and social platforms. The fact that 'more information is instantaneously available to us than

those born 100 years ago would have accessed in their lifetime' makes the ICTs that help citizens to process it every day a crucial part of their lives (Hawkins & McMahon, 2020). One of the assumptions about information is that people are merely using it through a 'service of delivering,' and 'you don't own it' (Cleveland, 1985: p. 187). A newspaper or a phone is a delivery service for the information. The modern environment where 'the conditions for communication and co-ordination' are changed poses new challenges for society (Päivärinta & Sæbø, 2006). A printed newspaper could not instantly change content in response to reader engagement. Unlike its digital media successors, it could do so only over period of time, in later issues or editions. The digital media environment, in which information flows are continuous and responsive to user engagement, shifts the power towards the creators of ICTs. The tools being developed are often used in ways consistent with their initial design and, therefore, entail inherent constraints, although their use can shift over time as users adapt functions beyond the intentions of the original design. For instance, platforms initially designed for personal networking are now widely used for political communication, market campaigning, and news dissemination, extending their function beyond original design intentions. While a printed newspaper was constrained by physical space, such as the number of pages in each issue, modern ICTs are constrained by attention, since their effectiveness depends on sustaining user engagement within limited spans of time. The amount and quality of knowledge that can be communicated through ICTs directly correlates to the capabilities of the software developed to deliver it, whether as part of the operating system or as an additional feature that structures how information is accessed and exchanged.

The issues described above come together in the theory of the information society. James Beniger, in his account of the 'Control Revolution,' advances a deliberately revisionist argument, stating that 'the transformation from Industrial to Information Society had been essentially completed by the late 1930s' (Beniger, 1986). Beniger's claim challenges more conventional accounts that locate the emergence of the information society in post-war developments in computing, telecommunications, or microprocessing technologies. At the same time, the concept of the information society itself is generally dated to later theoretical articulations, with its emergence commonly traced to the late 1950s (Lehtonen, 1988). Cawkell argues that Fritz Machlup, who first wrote about 'The Knowledge Industry' in 1962, should be credited with articulating the concept, as the idea of an information society was already implicit in his work (Duff, 2013; Cawkell, 1984).

For the purposes of this dissertation the components of Frank Webster's definition of information society is used to structure the discussion as it complements the exploration of the impact of the AI on democratic society from different angles. In Webster's 'Theories of the information society' it is pointed out that the construction of the information society 'seems so obvious' that it is used by many as if 'it is not necessary to clarify precisely what they mean by the concept' (Webster, 2014: p. 8). To counteract this passivity, Webster proposes to frame discussion of the information society using six different factors. This granular method helps to define in 'what ways and why information is becoming more central today' (p. 10). This thesis uses four components of Webster's definition to structure the analysis: technology, economy, culture, and quality.

The dimensions of employment and politics are set aside at this framing stage for methodological reasons. Employment is excluded because the analysis does not address labour markets, workplace organisation, or employment relations, but instead examines how AI-mediated systems shape participation through communication and discourse. Politics is excluded in Webster's institutional sense, that is, as the functioning of political parties, electoral systems, and formal governance structures. Rather than analysing political institutions directly, the thesis focuses on the discursive and cultural conditions that precede and shape democratic participation, including how RS structure visibility, relevance, and access to knowledge. Legislative and regulatory frameworks are therefore examined later in the thesis not as expressions of institutional politics per se, but as sites where these discursive conditions are articulated and contested through language. This approach allows the analysis to remain coherent with the initial framing while addressing regulation as a downstream manifestation of broader epistemic and communicative dynamics within democratic systems.

Technology

The first element of the definition that Webster uses is 'technological' (Webster, 2014: p. 11). The commonly predicted signs of the information society are 'new technologies' due to being 'one of the most visible indicators of new times' (p. 11). Webster refers to the 1970s and 1980s as periods during which it was claimed that 'that new technologies were of such consequence that they were thought to be bringing about systemic social change' (p. 11). This is similar to the current state of the AI discourse where researchers speak of the 'forthcoming AI revolution' and its effects (Makridakis, 2017). This thesis explores

practical examples of AI's impact with the notion of the 'systematic social change' (Webster, 2014: p.11) occurring. Webster states that 'computer communications currently inspire most speculation about a new society' driven by fact that the existence of such powerful and impactful technology is already a reason to identify a new form of society (p. 12). Webster's definition of information society suggests that modern technologies lead to 'a reconstitution of the social world because its impact is so profound' (Webster. 2002). There is an argument made by David Dickson that the society creates technology and is merely trying to 'respond by adjusting to the new' and, therefore, it is still the same society (Dickson, 1974). This position contrasts with Webster's claim that new technologies can reconstitute the social world, highlighting a tension between views that see technological change as transformative and those that frame it as an adaptation within existing societal structures. For this reason, the thesis adopts a multidimensional approach, drawing on Webster's definition of the information society to examine the effects on current and emerging democratic society operating in an environment already influenced by AI (Webster, 2014: p. 14).

AI is a metaphorical term used for a wide range of technologies. Numerous definitions of AI exist, but none are universally adopted. There are usually two types of definitions: a '*dictionary* definition is descriptive, as it summarizes the existing usage of the term' and a '*working* definition which 'is prescriptive, as it specifies a proposed usage of the term' (Wang, 2019). For the purposes of exploring how AI is defined, the definition developed by the Institute of Electrical and Electronics Engineers (IEEE) is particularly illustrative, because

it approaches AI as a layered system and divides its development into distinctive stages, showing how technical framings construct the term in contrast to broader descriptive or prescriptive definitions:

Artificial intelligence engineering has been compared to knowledge engineering. A 'knowledge-based system design' of AI encompasses 3 levels: 'the 'knowledge level' view of a knowledge-based system describes the knowledge that is used by and embedded in that system. The 'algorithm level' view describes the system as a search algorithm, configured out of standard component types (e.g., generators, testers, patchers, constraint propagators, belief revisers, etc). The 'program level' view expresses the system in terms of the elements of existing programming paradigms (rules, objects, procedures, etc) (IEEE, 2017: p. 10).

Not every instance of technology has the same impact on society. AI driven by machine learning (ML) in particular is so ubiquitous due to 'the availability of increasingly powerful computers, large and well-curated datasets, and developments in theoretical understanding of various learning algorithms' (Gündüz et al., 2019: p. 2184). The majority of contemporary ICTs, their operating systems and software, which include virtual assistance, device security, photo apps, writing aids and augmented reality apps, has a built-in AI of some sort (Biersdorfer, 2022; Marr, 2019). However, not all AI-enabled functions shape social and political life in the same way. Many applications operate at the level of optimisation, automation, or personal convenience, while others intervene directly in the organisation of information, attention, and visibility. ICTs such as phones, laptops, smart TVs, and tablets are particularly impactful on social life because 'mere transmission and reception of' information 'does not constitute communication' and therefore the quality of 'information content' is determined by affordances created by the software (Cherry, 1966: p.

9). Because these devices integrate AI into everyday practices of communication and interaction, they shape how people access, evaluate, and circulate information. This thesis therefore investigates the possible implications of such technologies for democracy by focusing on RS. This narrowing reflects the fact that contemporary forms of cultural participation and everyday discourse are largely mediated through online platforms that rely on RS to organise, prioritise, and circulate content. As a result, RS directly mediate access to information, structure visibility within digital environments, and shape the circulation of discourse in everyday communicative practices. For this reason, RS constitute a particularly suitable case for examining how AI-mediated technologies configure the conditions of participation relevant to democratic processes, while further technical detail is provided in the next chapter. While RS constitute a broad and diverse field, this dissertation does not assume that all RS are equally relevant to democratic processes. The analysis focuses on RS used by large-scale video platforms that curate culturally and politically relevant content and thereby shape visibility and access to information in everyday discourse. The empirical focus on YouTube and Dzen reflects their role as video-hosting platforms in the US and Russian contexts examined in Chapter 3.

Economy

Webster points out the importance of ‘the growth in economic worth of informational activities’ (Webster, 2014: p. 12). The increasing influence of the information business on the gross national product signifies ‘the achievement of an information economy’ (p. 15). The tendency of information economists ‘to put a price tag on everything has the unfortunate consequence’ of reducing

information to its market value and can lead to the failure to identify 'the really valuable dimensions of the information sector' that are not easily monetised (Webster, 2014: p. 16). Webster compares the 'the multimillion sales of *The Sun* newspaper' with the '400,000 circulation of the *Financial Times*' (p. 16). The economic gap between them is enormous. This is similar to the Matthew effect on YouTube, where more of the same information can lead the user 'to become more overconfident' (Tsai, Klayman & Hastie, 2008). Hence, the establishment of society where 'informational activity is of great weight but in terms of the springs of economic, social and political life is of little consequence' (Webster, 2014: p. 16). Webster creates a vivid image of economic information society as 'a nation of couch potatoes and Disney-style pleasure-seekers consuming images night and day' (p. 16). In this vein, this thesis explores the connection between the usefulness of the AI technology and its impact on the characteristics of information the person is able to operate with, which is influenced by economic aspects driven by stakeholder interest within technological environments (Van Caneghem & Van Campenhout, 2012). The economic gap between stakeholders, discussed above, is reflected in differentiated content environments within which RS operate, contributing to unequal conditions of exposure that are associated with a civic empowerment gap. This is evident in research on the civic empowerment gap, where different groups of citizens have unequal access to knowledge.

The civic empowerment gap, understood as the systematic inequality in opportunities for different social groups to engage in political and civic life, may be shaped by AI-mediated information environments that condition access to and

visibility of knowledge. The difference in civic participation of distinct socioeconomic groups is usually connected to their income (p. 10). The civic empowerment gap is based on the idea that 'poor people [or Hispanics, etc.] participate less. They don't have the time or financial resources (or education, knowledge, ...) to participate as wealthier people do,' which is considered 'inimical' to democracy (p. 10). As such, this dissertation recognises the need for an economic perspective on RS and, therefore, the literature review examines the cost function of the development process. The gap between citizens and government is escalated by the gap of technological accessibility. This issue occurs 'across all dimensions of good citizenship' (Levinson, 2010: p. 316). All of these factors condition citizens' capacity for civic participation by shaping access, visibility, and modes of engagement. Within this thesis, accountability and transparency are treated as enabling conditions for participation in AI-mediated environments rather than as guarantees of democratic outcomes.

Culture

Webster argues that 'an extraordinary increase in the information in social circulation' is easily noticeable (Webster, 2014: p. 21). This is especially the case due to the state of social media. The 'informational features of our world are more thoroughly penetrative than' ever and AI-driven technology enhances this environment (p. 22). Webster remarks that 'contemporary culture is manifestly more heavily information-laden than its predecessors' (p. 22). This nourishes the environment where 'signs lose their meaning and people simply take what they like from those they encounter' (p. 23). This leads people to mix 'different images to present no distinct meaning' (p. 23). Webster quotes Mark Poster to

summarise this definition and says that in this information society 'we have a set of meanings [which] is communicated [but which] have no meaning' (Poster, 1990 , p. 63). Later in this chapter, the implications of an information-saturated environment in which meanings become fragmented are discussed in relation to how conditions for political communication and interpretation may be shaped within a given information society.

The issues identified by Webster in the construction of the information society, and his observation that it 'seems so obvious,' can also be seen with the term democracy (Webster, 2014: p. 8). Democracy 'is one of the most used and contested concepts in social science' (Baviskar & Malone, 2004: p. 4). Perceptions of democracy vary due to its complex nature. The desire and motivation to participate in a democratic society stem from individual understanding, perception, and associations with the term. Words 'are empirical signs,' and their meaning depends 'upon the listener, upon the speaker, upon their entire experience of the language, upon their knowledge of one another, and upon the whole situation' (Cherry, 1966: p. 11). In this sense, democracy is grounded in shared values, including equality, diversity, equity, transparency, and freedom (Diamond, 1999). The ways these values are interpreted and prioritised vary considerably across individuals and groups, shaping different understandings of what democracy entails in practice.

The complexity of the concept of democracy divides the definitions used in democratic theory into three types: *minimal*, *middle-range*, and *maximal*, which share a common ground in terms of civic participation (Schlenkrich, 2019). The

scientific community generally agrees on a *minimal* definition – ‘the repeated holding of elections with competition and broad participation’ (p. 316). The *maximal* definitions can be seen to ‘overstretch the concept of democracy’ (p. 316) and ‘conflate democracy with other concepts that would be used to explain democracy or that democracy would be used to explain,’ but have a useful participation element as a foundation in them (Davenport & Armstrong, 2004). Middle-range definitions replace ‘vaguely defined entities with more concrete, circumscribed concepts, and substitute hypotheses about specific relationships between those concepts for grander theoretical frameworks’ (Mutz, 2008: p. 522). In practical terms, this approach shifts analysis away from abstract claims about democracy as a whole towards identifiable mechanisms, such as particular forms of discourse, information exposure, or participation practices, and their observable effects. Rather than treating democracy as a single, unified outcome, middle-range definitions allow specific elements of democratic life to be examined in isolation, making it possible to assess how distinct communicative conditions contribute to or constrain civic participation.

One of the critical components that function beneath the umbrella term democracy is civic participation. While it can be defined as the voluntary act by citizens who ‘in a democracy seek to control who will hold public office and to influence what the government does’ (Schlozman, Verba & Brady, 1999: p. 430), it also encompasses broader forms of engagement. These include community organising, public deliberation, cultural participation, and digital practices that contribute to shaping collective life. Civic participation is important because it encourages citizens to educate themselves, benefiting individuals, communities,

and democracy as a whole. Civic participation also has an impact on elected officials and informs them about the state of general public, which helps them govern in more inclusive way. The quality of inclusivity is investigated further in the next chapter when statistical discrimination and data provenance in RS is discussed.

Quality

This dimension of Webster's definition is the most crucial for this dissertation. In contrast with the three definitions above, this one does not argue about the increased amount of information. Webster claims, 'that the character of information is such as to have transformed how we live' and proposes to define the information society by a quality of information and not by quantitative approaches (Webster, 2014: p. 11). Webster talks about Herbert Schiller's 'garbage information' and compares it to 'junk food' (p. 32), using also Susan Jacoby's (Jacoby, 2009) formulation regarding the 'amnesia of audiences locked into an internet loop' to make a point about the quality of information (Webster, 2014: p. 32). This thesis examines the contemporary environment in which AI influences what, how, and where people access information, and how this information is transformed into knowledge. The emphasis is on RS output quality and the linguistic diversity of recommended content, since these factors affect whether information can be meaningfully processed, interpreted, and integrated into democratic discourse. The analysis considers whether individuals have the necessary tools to acquire and analyse the recommended content. If 'citizens are equipped with the skills needed to be data literate' in this 'data-driven society' (Wolff et al., 2016), the impact of RS on democratic processes would be

mediated by their ability to critically evaluate recommendations. Yet the extent to which citizens actually possess such skills is contested, and without widespread data literacy individuals remain vulnerable to the influence of RS despite the availability of information.

The New York Times article discussed earlier illustrates such an impact. Kevin Robins and Frank Webster discuss how the forms of information and data-driven society nourish social control patterns (Robins & Webster, 2003: p. 180), emphasising the sophistication of military ICTs and surveillance, and arguing for 'the centrality of the military paradigm within the current process of economic and social restructuring' (p. 162). They also argue that the quality of education declined as a result of programmed standardisation, which they see as part of a broader effort to create a more controllable society. This focus on education is significant here because it links the structuring of information environments to citizens' capacity for critical engagement, a theme central to understanding how ICTs and RS influence democratic participation. Chapter 3 develops this point further by comparing the amount of information present in RS output with the vocabulary size required for users to comprehend it. Edward Alsworth Ross states that 'our institutions are not shaped by any one simple ethical principle that appeals to all men who are not bad' (Ross & Gross, 2017). The 'hopeless contradictions' are inevitable in the process of search and analysis (p.55). According to Robins and Webster, social control patterns formed in education and fed by a military information society created 'education for flexibility,' making society more susceptible to certain forms of mediated messaging (Robins & Webster, 2003: p. 171). This focus on education is significant here because it

links the structuring of information environments to citizens' capacity for critical engagement, a theme relevant to examining how ICTs and RS condition exposure within civic communication. Formal education is considered to be the most impactful factor increasing political participation, as it equips citizens with the cognitive and linguistic resources needed to engage with complex political discourse (Glaeser, Ponzetto & Shleifer, 2007). Similarly to the civic empowerment gap and its relationship to education, a person's ability to comprehend information depends not only on their educational background but also on the environment shaped by ICTs, which influences how information is presented and accessed (Levinson, 2013).

AI-driven technologies structure information environments in which phenomena such as the 'paradox of choice,' infobesity, and information overload may emerge, shaping the conditions under which information is encountered and processed (Bawden & Robinson, 2009). The 'establishment of printing in Europe was accompanied by criticisms of the quality of the material being produced, which evokes a striking resonance with recent views of the quality of web resources' (p. 182). With the quantity and quality of information having 'opposing influences on decision effectiveness,' AI-driven ICTs 'overload' citizen with the information, which leads to 'a decrease in decision effectiveness' (Keller & Staelin, 1987). Webster's definition of an information society focusing on the quality of content discusses the possible implications of such environment. Edward Bernays argued that the organisation of society 'by leadership and propaganda' is 'necessary to orderly life' (Bernays, 1928: p. 12). Modern media take the information and converts it 'into a partisan package delivered to its

various constituents: a steady flow of bias- confirming stories that create a shared narrative of the state of the world’ to keep the order (Benkler, Faris & Roberts, 2018). At the same time these outlets ‘are not designed to check or refute propaganda as long as it is consistent with the partisan narrative’ (p. 80). The ‘negative responses to a media statement or program’ that were known as ‘flak’ (Herman & Chomsky, 2010: p. 26) are not needed anymore due to the large amounts of information. An RS’s workflow nourishes this narrative-based environment and is studied in Chapter 3 as seen operating on YouTube and Dzen.

The West versus Russia

To be able to better demonstrate the nature of how an algorithm-driven system interacts with democratic and civic participation, this thesis explores the influence of technology on two contrasting political contexts: liberal-democratic systems represented by the United States and authoritarian governance represented by Russia. The reference to “the West (including North America and Europe)” is used descriptively to indicate broadly comparable liberal-democratic settings rather than to imply political or cultural uniformity across different national contexts.

The West in the nineteenth century went through ‘a period of fundamental political reform and unprecedented changes in taxation and redistribution’ (Acemoglu & Robinson, 2000), which created an environment in which liberal-democratic institutions became central reference points in international politics (Biden Jr & Carpenter, 2018). The end of ‘the East-West conflict [Cold War]...coincides with a remarkably broad validation of free elections, democratic

politics, and market economies' (Farer, 1993). The narrative about Russia in recent decades has changed from the Cold War perspective of the USSR into one of 'a country many knew was becoming more difficult but a country few believed would actually resort to military force to attack one of its neighbours' (Asmus, 2010: p. 215). This narrative was proven wrong in 2008 during the war in Georgia and again in 2014 and 2022 with the war in Ukraine. In 2007 *The New York Times* editorial stated ahead of Russian parliamentary elections that 'this is no free and fair election and Russia is no democracy' (The New York Times Editorial, 2007). Russia is widely recognised as authoritarian, with international indices documenting persistent limits on political competition, repression of independent media, and restrictions on civil society (Tsygankov & Parker, 2015; House, 2018, 2017; Gorokhovskaia, Shahbaz & Slipowitz, 2023; Unit, 2024; Levitsky & Way, 2010). Nevertheless, both Russia and states such as those in the EU and the USA describe their systems as democracies. In this dissertation, Russia is treated as an authoritarian system that nonetheless claims democratic status, a distinction that is analytically relevant for examining how RS are framed, regulated, and deployed.

An important context for understanding the role of AI in shaping communication between individuals in these settings is how emerging communication norms affect democratic exchanges. In this thesis, democracy is approached primarily through the lens of civic participation and discourse rather than as a comprehensive regime classification. Democracy requires rational discussion about what is best for the community (Berelson, 1952: p. 322). This assumption reflects 'Aristotle's view' of debate in legislatures as a 'democratic

form of argument' (Keith, 2007: p. 126). To examine the broader impact of technology on the democratic process, Chapter 3 considers a case where AI disrupts civic engagement through cultural participation on video-hosting platforms that use RS in the US and Russia.

How RS diminish Democratic Processes

AI affects the primary communication tool people use, language. The ability to communicate and interact depends on developing language skills, which are important for civic and cultural participation. Control over information and communication systems creates a linguistic diversity gap, defined as a reduction of variety in language and perspectives available in recommended content due to the data processing and structuring functions of the technology (Viswanath & Finnegan, 1996: p. 209). Algorithms pose challenges of statistical discrimination and the need for concepts shaping AI's flow to be 'static enough to be encoded' (Wachter, Mittelstadt & Russell, 2021). When AI uses language as a dataset, its mathematical design omits some data in favour of speed and efficiency. This dissertation situates the cause of such omissions in the educational and developmental aspects of AI, using these terms in a technical sense to describe how systems are trained on data and refined through iterative learning processes rather than in relation to social or human development. Building on this, this thesis assesses civic participation and the epistemic agency gap shaped by AI-driven technologies against key democratic principles. In this context, the gap signifies a constraint, as it reduces rather than expands citizens' capacity to participate effectively. The term 'epistemic agency' refers to the ability of individuals to play an active role in how knowledge is produced, assessed, and

shared, rather than only receiving it. Epistemic agency has been defined as ‘the power to shape the knowledge production and practices of a community’ (Miller et al., 2018: p. 2). In philosophy, virtue epistemology approaches knowledge by focusing on the qualities and abilities of the knower (Greco & Turri, 2017; Battaly, 2008; Axtell, 1997). Virtue epistemology treats knowing as arising from good intellectual habits, such as being fair-minded, careful, and willing to consider evidence, as well as from using reliable ways of finding out information, such as accurate observation and memory. In virtue epistemology, epistemic agents are described as ‘legislating members of a realm of epistemic ends’ who establish and follow the standards, methods, and rules that govern epistemic practice (Elgin, 2013: p. 136). Dotson’s systems approach examines structural limits on epistemic agency, identifying first-order exclusions, such as testimonial injustice, second-order exclusions, such as hermeneutical gaps, and third-order exclusions, where a system is unsuited to the interests of certain knowers (Pohlhaus, 2020: pp. 2-4). Inclusion can also be ‘epistemically exploitative,’ extracting epistemic labour in coercive or non-reciprocal ways (Pohlhaus, 2020: p. 2). For instance, individuals may be formally included in participatory or discursive processes while being expected to contribute experiential knowledge, linguistic labour, or data without having meaningful influence over how that knowledge is interpreted, valued, or used. In such cases, inclusion increases visibility or data extraction without corresponding epistemic authority or agency, resulting in participation that benefits the system more than the participant. Debates on scope range from mental activities such as judging or revising beliefs to arguments that belief is a static state, even if held ‘on the ground’ of other beliefs (Engel, 2013; Setiya, 2013). Technology can embed structural exclusions

into its design, shaping attention and constraining what can be known (Pohlhaus, 2020: p. 4). RS guide access to data for citizens. The same applies to language, where AI reduces opportunities to practise and use diverse linguistic resources, creating a linguistic diversity gap. This reduction limits citizens' ability to shape the knowledge they access and share, producing the epistemic agency gap. When individuals have less control over how they participate in knowledge production, differences in their capacity to engage are amplified, contributing to the civic empowerment gap. The relationships between linguistic diversity, epistemic agency, and civic participation are treated in this thesis as interrelated constraints rather than as a causal sequence. Reductions in the range of available expressions can limit the resources individuals draw on when interpreting and shaping knowledge, which in turn conditions opportunities for participation within civic communication environments. This narrowing in turn constrains epistemic agency, as individuals have fewer resources to shape and evaluate the knowledge they encounter. The diminished agency then translates into unequal opportunities for civic participation, reinforcing the civic empowerment gap. Democracy depends on the effective and collaborative operation of diverse knowledge- and information-sharing mechanisms. A core principle for productive citizen participation is access to knowledge for everyone who needs it (Tandon et al., 2016: p. 22).

Although many applications of AI disrupt social and cultural processes, certain types of tools merit particular attention for their potential impact on democracy. This dissertation focuses on RS, which 'aggregates and directs to appropriate recipients' the data provided by users (Resnick & Varian, 1997).

Because these systems implement behavioural data as a core input, they can be understood as prominent examples of ML based on user engagement and participation (Farzan & Brusilovsky, 2011). They operate on behavioural data as a core input and structure information environments in ways that may increase cognitive load and and ‘a decrease in [a person’s] decision effectiveness,’ without establishing direct effects on individual behaviour or levels of civic participation. (Keller & Staelin, 1987). The accumulation of citizens’ input is vital for a healthy democratic environment (Claassen, 2020), and exclusion from this process leads to marginalisation. For instance, studies show that people from China and India are underrepresented in most datasets used for ML and that some algorithms have been characterised as ‘racist’ and ‘sexist’ (Shankar et al., 2017; Zou & Schiebinger, 2018; Massanari, 2017; Chander, 2016). These dynamics are particularly visible with technologies that personalise and filter information.

The impact of this marginalisation is not limited to social exclusion but extends to a reduction in human agency, which is shaped by the AI-driven systems integrated into decision-making processes. The ability to choose from different alternatives is a basic cognitive process (Wang & Ruhe, 2007: p. 73). Modern RS mostly use ‘collaborative recommendation’ techniques (Burke, Felfernig & Göker, 2011: p. 15; Schafer et al., 2007), an approach that blurs individual experiences and narrows the range of choices presented to the user. This method presents the targeted person with preferences expressed by others with the same initial choice (O’Mahony et al., 2004: p. 344). Such curation of human–technology interaction reduces the user’s capacity to act as an independent epistemic agent. In cases where systems ‘redirect decision makers

without their explicit permission,' questions arise about the 'free will paradigm' and the implications of neuromarketing strategies driving the development of RS (Wilson, Gaines & Hill, 2008). This frames democratic participation as occurring within information environments that may be shaped by rapidly developing technological tools whose impacts are difficult to anticipate, rather than as a direct effect on democracy itself. In spite of the perceived 'demand for ethically and socially' aligned data processing there is almost no 'regulatory framework to address these issues' (Mantelero, 2018: p. 755). The concepts of accountability and transparency of AI are usually formulated through examples 'when such systems cause harm' (IEEE, 2019: p. 5), but sometimes the harm is not visible, it is an opportunity or hidden cost. The risks of these hidden costs are difficult to manage, since policymaking for technological communication tools requires dialogue 'with stakeholders of diverse backgrounds' and often fails to reach a conclusion (Aizenberg & Hoven, 2020: p. 10). These issues are discussed in the course of this dissertation in the context of existing laws using the EU AI Act as a case study in Chapter 4 (European Parliament, 2024). In 'The Tragedy Of The AI Commons' Travis Lacroix and Aydin Mohseni used stochastic evolutionary game dynamics to predict if the recommendation report on AI created by the United Nations Educational, Scientific and Cultural Organization (UNESCO) would be at all effective compared to other reports 'containing ethics principles for AI research' (LaCroix & Mohseni, 2021: p. 1). The conclusion was that 'there may be little reason to think that this document will be much more effective than any of the other 84 documents that have been proposed by NGOs, governments, corporations, professional associations, etc.' (p. 17). This illustrates the difficulty of turning ethical principles and policy proposals into effective regulatory

mechanisms, highlighting the limits of soft governance in addressing the risks posed by RS and other AI-driven technologies. LaCroix and Mohseni proposed ‘cooperation that spans international borders’ as a possible solution (p.18). The literature frames international policy discourse as a proposed mechanism for coordinating responses to AI-related risks, while also documenting persistent limitations in its practical effectiveness, a tension examined in Chapter 5.

One key factor undermining such cooperation and the development of shared standards is the persistence of biases embedded in AI communication design, which lead to creation of tools with environments where sexism, racism, ageism, ableism, and other biases are amplified. Decision-making processes are undermined by unconscious biases (Korte, 2003: p. 440). The view of technology as a ‘magnifier of underlying human and institutional intent and capacity’ suggests that it reflects both positive and negative aspects of human life (Crandell & Smaldino, 2000). These biases are further shaped by the ‘problem formulation’ choices of developers, which are influenced by culture and programming language paradigms (Srinivasan & Chander, 2021: p. 46). This thesis examines the real-world implications for language shaped by ML outputs by analysing the language used to build RS and the limitations embedded in them.

Questions about the possibility of AI influencing political systems, and democratic principles in particular, remain a topic of debate (McGinnis, 2012: p. 193; Glassman, 2019: p. 189). Two issues identified in this context are the decentralisation of governmental regulation systems and the status of human

rights. The first considers AI as a dissemination tool that can expand access to knowledge, potentially disrupting ‘top-down decision making’ and enhancing human agency (McGinnis, 2012: p. 191). Decentralised systems are recognised for their ability to reflect the ‘accountability of governments to their citizens even without differences in preferences between localities’ (Seabright, 1996). The second issue addresses debates around the ethical and legal implications of environments in which AI is discussed in relation to human rights, raising questions about accountability, autonomy, and legal personhood (Solum, 2020; Lau, 2019; Brown, 2021). While these debates are relevant to broader discussions of AI governance, they are not pursued further in this dissertation, which focuses instead on how AI systems, and the environment created by RS in particular, affect epistemic agency and the conditions of democratic participation.

Understanding these political and ethical concerns requires examining the technical foundations of RS, as the design choices that shape them determine how information is filtered, prioritised, and presented to citizens. This thesis examines AI based on ML, as other forms of AI are less common and more industry-specific, and therefore have less impact than ML applied to textual data. Different ML implementations have distinct methods of data processing. A common practice in computer science is to adapt algorithms built and tested on one type of information, such as images, for use on other types of datasets. Chapter 2 examines the most ethically vulnerable stages of RS development to understand how AI processes data. The efficacy of ML depends on the efficiency of data structuring and organisation. Common practices for restructuring

datasets to make them computationally efficient often prioritise one body of knowledge over another, reducing knowledge diversity and constraining citizens' ability to exercise their democratic rights.

Approach of this Thesis

The core aim of this dissertation is to demonstrate the potential threat AI poses to democracy, specifically how algorithmic filtering, particularly through RS, reduces linguistic diversity and expressive range, limiting citizens' capacity to engage in rational, informed, and constructive democratic processes. Its primary research questions are:

1. Whether these limitations and their impact operate differently in the United States, which primarily uses YouTube, and in the Russian equivalent, Dzen;
2. How regulatory approaches in Europe, the United States, and Russia address the harmful effects of RS, and how these approaches reveal implicit interpretations of their interaction with democracy and civic engagement;
3. How the language used in RS discourse in English and Russian, and the ways key terms are defined and operationalised, reflect differing conceptualisations of RS and their democratic implications.

This dissertation examines the influence of AI-driven technologies on the foundational elements of democracy and their implementation in everyday life. It does so through three stages in which democratic principles and the affordances and constraints of AI intersect. The first stage analyses YouTube and Dzen,

focusing on the level of political polarisation sustained by each platform's RS. In the United States, the study considers the capacity of RS to create 'rabbit holes' that produce information homogeneity and reduce decision effectiveness. In Russia, it examines algorithms and approaches in RS that may affect the linguistic variation available to citizens discussing the war in Ukraine, thereby influencing their ability to engage meaningfully in democratic processes. The second stage examines recent AI policy documents, focusing on policies in the EU, US, and Russia, to assess how they address the risks and constraints identified in the first stage. The discourse of these policy documents is analysed for similarities and differences in the way the perspectives of three major stakeholders, software developers, regulators, and users, are represented. The third stage investigates RS discourse, examining how software developers, regulators, and users describe and characterise RS. Unlike the second stage, which focuses on policy texts, this stage analyses the language used within and across these groups. It identifies differences in discussions not only between fields but also within disciplines, highlighting discrepancies among representatives of the same field. This progression moves from platform case studies to policy analysis and then to disciplinary discourse, enabling a broader examination of the challenges AI-driven technologies pose to society. The findings from the first stage are compared with their representation in AI policy and RS discourse. These approaches are linked by the role of communication and language, with changes in discourse forming an overarching theme across the analysis of AI's impact on democracy, regulatory responses, and stakeholder narratives.

Methodologically, this staged approach is designed to trace how RS shape democratic participation across interconnected levels of analysis. Beginning with platform-level case studies allows the thesis to identify concrete effects of algorithmic curation on discourse and participation. Examining policy documents in the second stage situates these effects within institutional responses and regulatory framings, while the third stage analyses how these dynamics are articulated and normalised within research and professional discourse. Taken together, the three stages allow the analysis to move from observable platform practices to regulatory interpretation and finally to conceptual framing. This cumulative structure enables Chapter 6 to synthesise the findings across all stages, positioning RS within a broader configuration of cultural-technical systems that shape communicative conditions and democratic participation.

Structure of this Thesis

Chapter 2 presents a literature review focussing on a specific implementation of ML — RS. It examines the current state of the art in the definition, implementation, creation, and risks of RS. The aim is to investigate the processes underlying this type of AI, with particular attention paid to the workflow of these systems in relation to datasets. The capacity of this technology to ‘decrease information diversity’ (Bozdag & Van Den Hoven, 2015) and shape the ‘individual experience of digital environments and social interactions’ (Milano, Taddeo & Floridi, 2020) affects citizens’ ability to participate rationally in democratic processes. The outcome of this chapter is a structured account of all stages of RS and the identification of potential threats to democracy through the

algorithmic narrowing of content exposure, commonly described as a ‘filter bubble’ (Pariser, 2011; Nguyen et al., 2014).

Chapter 3 examines how RS shape linguistic diversity and political discourse in different democratic and authoritarian contexts. It draws on two case studies to compare the impact of RS-driven content curation on YouTube in the United States and on YouTube and Dzen in Russia. The US case study analyses content recommended during the four months preceding ‘Super Tuesday’ 2024, a decisive stage in the presidential primary cycle (Pathe, Krieg & Bradner, 2024). The Russian case study focuses on political discourse during the four months preceding the 2024 Russian presidential election, when the war in Ukraine dominated both state-aligned and oppositional narratives. By comparing opposing viewpoints in both contexts, this chapter demonstrates that RS technologies can reinforce polarisation, narrow linguistic exposure, and constrain democratic participation, even when operating within divergent political systems.

Chapter 4 examines how the effects of RS technologies identified in Chapters 2 and 3 are recognised and addressed in AI policy frameworks proposed in the United States, the European Union, and the Russian Federation. It analyses the 2024 draft of the EU AI Act (European Parliament, 2024), the 2023 Executive Order (EO) of the President of the United States (The White House, 2023), and the 2024 EO of the President of the Russian Federation with the accompanying National AI Strategy (Kremlin, 2024). The analysis compares how each framework addresses RS-specific challenges, including cold start, algorithmic bias, echo chambers, and overfitting. The chapter incorporates

cultural analysis using Hofstede's Cultural Dimensions Theory, Schwartz's Theory of Basic Human Values, and Fiske's Relational Models Theory to interpret differences in governance discourse (Hofstede, 1984; Schwartz, 1992; Fiske, 1991). The findings are evaluated in terms of their implications for epistemic agency gap and are compared with the results from Chapters 2 and 3.

Chapter 5 examines how RS research discourse frames regulation, agency, risk, and technical function. Drawing on RS research in English and Russian, it analyses how terminological patterns affect epistemic accessibility of knowledge about recommendation technologies. The analysis focuses on RS researchers' own language and tracing how institutional assumptions are embedded through shifts in terminology and framing. Moving beyond the Technology, Economy, Culture, and Quality categories established earlier, the chapter adopts four themes: governance and accountability, agency and autonomy, technical instrumentalism, and risk and mitigation. These clusters capture how research discourse constructs its subject through lexical emphasis and omission. Each section combines conceptual framing, frequency analysis, network metrics, representative excerpts, and glossary comparisons to assess how RS research contributes to epistemic containment and shapes public intelligibility of AI systems. Comparative analysis of English- and Russian-language corpora highlights differences in technical framing, governance alignment, and cultural assumptions.

Chapter 6 synthesises the findings of the preceding chapters to assess how RS influence democratic processes through their impact on linguistic

diversity, epistemic agency, and regulatory discourse. It integrates empirical results from the case studies in Chapter 3, the policy analysis in Chapter 4, and the research discourse analysis in Chapter 5 to trace a progression from micro-level algorithmic effects to macro-level institutional framing. The chapter evaluates how a five-perspective chain, spanning technical design, economic incentives, cultural framing, content quality, and governance, converges to produce managed plurality in online communication. This managed plurality, while allowing limited diversity, constrains interpretive range and reduces citizens' epistemic agency. Linking the experimental, regulatory, and discursive dimensions, the chapter positions RS as both a technological and cultural actor that shapes the boundaries of democratic participation.

Chapter 7 presents a conclusion where the findings of Chapters 3–6 are combined to develop a unified perspective on the gaps and disconnections in understanding the impact of RS on democracy. It returns to the concept of civic participation and analyses it through the structure of the definition of information society discussed above. The goal is to provide a stable understanding of the weak links in the modern information society that can threaten democracy, using RS as an example of AI.

Chapter 2. Literature review

This chapter examines the state of development of recommender systems (RS), focusing on their design, deployment, and evaluation, and considers software development practices that can embed ethical risks into these systems. These practices are often standardised within industry workflows and validated through performance benchmarks, yet they are rarely subject to sustained critical examination outside academic or policy settings. Littman et al. note that RS ‘have become absolutely essential’ and are now widely used, yet the literature also identifies persistent concerns about ‘fairness, diversity, polarisation, and the emergence of filter bubbles’ (Littman et al., 2022). While ‘much of the [academic] work [on AI ethics] has focused on formalising anti-discrimination in the context of classification systems’ (Mitchell et al., 2020), these systems depend on fixed categorisations of data and attributes, which may not capture the complexity of social contexts. In the case of RS, ethical considerations extend to privacy, personalisation, and the narrowing of informational exposure through feedback loops.

This chapter follows the structure outlined in the introduction to this dissertation, drawing on four of Webster’s definitions of the information society, highlighting aspects of technology, economy, culture, and quality (Webster, 2014). These headings provide a progression from identifying the state-of-the-art in RS to examining the methods used to measure quality during development. In later chapters, this review of RS development practices is revisited to illustrate how embedded assumptions, and conceptual limitations may be transferred into the frameworks that researchers and policymakers rely on when addressing

ethical challenges in AI. The analysis draws on literature about software creation to examine RS development through the lens of developers as ‘human decision makers [who] trust more in the accuracy of the recommendations of the algorithms’ than in their own capacity to address data-processing concerns, and considers the implications of this tendency (Coeckelbergh, 2020: p. 130). Such development processes can embed assumptions that, under certain conditions, are associated with the persistence of biases and with professional environments in which alternative perspectives may have limited influence. The ethical concerns discussed here are linked not to ‘statistical or mathematical knowledge as such, but [to] the way it [statistical reasoning] becomes a default in’ RS (Coeckelbergh, 2022: p. 7). The chapter’s structure supports an exploration of how RS algorithms may condition aspects of epistemic agency within mediated information environments, without presuming direct or uniform effects on citizens (p. 6).

The *Technology* section outlines the historical development, main trends, and current implementations of RS, focusing in particular on collaborative filtering and content-based approaches, which remain the most widely used. These definitions provide a basis for understanding the range of stakeholders and overarching objectives associated with RS, treating them here as technological artefacts designed to address specific computational problems. Ethical risks are identified only at a broad level in this section and revisited later in the chapter, where algorithmic practices are examined in relation to RS cultural impact. The *Economy* section considers how developers evaluate the efficiency and productivity of RS, noting that the systems discussed in this thesis have been

developed in commercially driven contexts. This section examines the stages of development that directly affect the economic viability of RS, where computational approaches are prioritised according to their projected business impact. At this point, development choices shape the selection of programming practices and mathematical methods for addressing computational challenges, some of which emerge during the neural network training phase. Alongside a general review of evaluation and accuracy metrics, the section highlights the loss function algorithm, a common tool for training and optimisation. The ethical complexity of RS development is heightened by the fact that ‘the very basic problem is the appropriate choice of the loss function,’ which is defined on a case-by-case basis (Akaike, 1974). Later chapters, particularly Chapter 3, revisit the implications of this choice, examining how loss function design influences recommendation patterns on video-hosting platforms and how these patterns condition the informational environments within which cultural participation takes place. The *Culture* section examines data provenance in RS, focusing on ethical questions about the datasets used to train and refine these systems. Because data can carry multiple cultural layers, its handling presents distinctive ethical risks. Training datasets embed diversity, context, and bias into recommended outputs, yet the processes by which these layers are combined and transformed often lack oversight. This can result in certain perspectives being amplified while others are excluded. RS decision-making is sensitive to subtle patterns in the data, and the emergent cultural landscape within the neural network may reflect the prioritisation of mathematical problem-solving over attention to data sensitivity. This dynamic can create tensions with principles of inclusion by shaping which perspectives are more visible within mediated information

environments. The culture-related risks identified here connect directly to the analysis in Chapter 6, which considers how RS design choices, optimisation priorities, and governance contexts influence cultural participation. The *Quality* section addresses shrinkage effects, underfitting and overfitting, statistical discrimination, and differences between fast and long decision processes. These computational approaches shape the ‘quality of the software’ (Numan, 2020: p. 134), but they also define how RS prioritise and transform data. At the scale of modern AI, the ethical risks linked to these capabilities are amplified. The section compares contemporary RS objectives with the methods used to achieve them, setting the stage for subsequent chapters to examine how developer decision-making shapes RS design choices and the information environments within which civic participation takes place.

Technology: Historical development and dominant approaches in RS

Contemporary RS are complex systems embedded across many industries, where personalised suggestions and the analysis of customer preferences and behaviour underpin business strategy. RS provide ‘suggestions for items to be of use to a user’ in ‘decision making processes such as what items to buy, what music to listen [to], what online news to read or what learning objects to learn’ (Ricci, Rokach & Shapira, 2010). The same technology can structure how individuals encounter and navigate information that is relevant to political communication. Most digital and online platforms integrate RS to organise interaction and maintain engagement, though their scope and intrusiveness differ across contexts. Their effectiveness is often linked to the capacity to process large volumes of information and to increase reported user satisfaction. RS are

described as helping people ‘to deal with information overload’ by offering ‘personalized recommendations, content, and services to them’ (Adomavicius & Tuzhilin, 2005). In this framing, RS may function as tools that support individual decision-making in everyday and professional settings, with implications for how information is accessed and evaluated within civic communication contexts.

Early RS development centred on information filtering, primarily through collaborative filtering and content-based approaches designed to enhance the user’s personal experience. These methods typically relied on user–item interactions and keyword analysis, operating without semantic understanding and facing well-documented limitations such as the cold start problem, discussed later in this chapter. Collaborative filtering ‘helps people make choices based on the opinions of other people,’ supporting efforts to cope ‘with information overload and low signal to noise ratio’ (Resnick et al., 1994). In this model, RS ‘use the opinions of a community of users to help individuals in that community more effectively identify content of interest from a potentially overwhelming set of choices’ without providing contextualisation or explanation (Herlocker et al., 2004). These approaches became increasingly important in the 1990s as internet use expanded, and each implementation brought refinements in scale and efficiency. Their adoption marked a significant shift from earlier rule-based and popularity-based filtering, which offered limited personalisation and depended on aggregate trends. In one of the earliest implementations of collaborative filtering, systems recorded users’ reactions to documents they read to perform filtering (Goldberg et al., 1992). Content-based filtering, by contrast, recommended ‘best-matching items’ by comparing candidate items with those previously rated by the

user (Thorat, Goudar & Barve, 2015). Initially, both collaborative and content-based systems found applications in e-commerce, where their primary role was to match products to likely buyers. They were not developed with the explicit aim of managing sensitive information, nor with consideration of how their outputs might be taken up in contexts relevant to civic or political communication. Nevertheless, the widespread adoption of these techniques established the technical foundations for later RS innovations and expanded their role beyond commerce into domains with broader cultural and political significance.

Recent developments in RS include hybrid systems, deep learning techniques, and context-aware architectures, all of which have extended the scope and application of RS technology. Hybrid systems combine collaborative filtering with content-based methods 'to achieve peak performance' and leverage their 'complementary advantages and disadvantages' (Burke, 2002). This integration can address limitations such as the cold start problem, where a lack of user or item data constrains early recommendations, and can mitigate the homogeneity often associated with collaborative filtering. These gains, however, come with increased system complexity and scalability challenges. Deep learning-based RS process raw and high-dimensional data to model complex patterns and improve predictive accuracy (Zhang et al., 2019). Context-aware RS expand on traditional approaches by incorporating additional information such as time, location, or inferred emotional state to produce recommendations that are more closely tailored to situational factors (Adomavicius & Tuzhilin, 2010). Modern RS are described as 'critical tools to enhance user experience and promote sales/services for many online websites and mobile applications'

(Zhang et al., 2019). Their influence, however, extends beyond commercial performance. RS shape how users encounter and interpret information across multiple domains. These include news recommendation, political video platforms, music and cultural discovery systems, and educational or informational content feeds. In such settings, ranking and selection practices structure visibility, relevance, and exposure.

To assess the performance of advanced RS, researchers apply a range of accuracy metrics, including Root Mean Square Error (RMSE), precision, recall, and F1-score. Accuracy is measured by comparing the system's recommendations with recorded user preferences in the dataset and often becomes a central development goal, given its reported 'larger effect on user satisfaction and performance' (Herlocker et al., 2004). When accuracy is prioritised, and an RS aims to predict 'by considering the association between the item ratings and the corresponding item attributes' (Ricci, Rokach & Shapira, 2010), the narrowing of presented options becomes particularly pronounced, as recommendations, particularly in news and opinion-oriented contexts, can converge into echo chambers or epistemic bubbles shaped by opaque algorithmic processes rather than explicit user choices (Nguyen, 2020; Pariser, 2011). An RS trained to recommend items closely resembling those previously selected can generate a feedback loop in which the highest accuracy corresponds to increasingly limited decision pathways. While accuracy remains a key metric, it is typically supplemented by measures such as diversity, novelty, and user satisfaction. Diversity and novelty are intended to broaden exposure and reduce the likelihood of closed information loops, thereby supporting more

informed decision-making in complex information environments (Ziegler et al., 2005). Yet these measures face constraints. Selective exposure, in which users seek information that aligns with pre-existing beliefs and avoid contradictory perspectives (Nguyen, 2020), and the tendency to form opinions through personal networks (Kahan et al., 2012) can limit their effectiveness. These behavioural factors complicate the assumption that technical adjustments alone can counteract structural biases in recommendation outputs. In particular, user preferences for affirmation rather than informational challenge mean that recommendations which align with existing views are more likely to be selected, engaged with, and reinforced through feedback signals. Even when diversity or novelty metrics are introduced, systems remain constrained by interaction data that rewards confirmatory content over dissonant information. As a result, technical interventions operate within behavioural limits, where optimisation continues to privilege affirmation because it produces measurable engagement, making structural bias resistant to purely technical correction.

RS development is also shaped by considerations of scalability and efficiency. These factors are central when designing algorithms capable of delivering high-quality recommendations while sustaining the performance required for large-scale datasets and real-time processing. When a system fails to meet these requirements, such as during rapid user growth, it can produce ‘undesirable effects such as bias toward certain items, lack of diversity or imbalance in recommendations, and even potential security concerns such as making it easier to deliberately manipulate the system outcomes’ (Eskandarian, Sonboli & Mobasher, 2019). Alongside scalability and efficiency, user satisfaction

and engagement in commercial settings are often prioritised. These objectives focus on refining the user experience and supporting long-term retention. User satisfaction is shaped by 'subjective experience data' and is influenced by the extent of human agency in the interaction (Knijnenburg et al., 2012). To sustain engagement patterns, 'behavioral data should therefore be complemented with subjective experience measurements' (p. 452). In commercial settings, RS frequently measure satisfaction through user feedback and retention data, using these metrics 'as a leverage for holding the user tightly to the application,' reflecting 'service provider's motivations' (Ricci, Rokach & Shapira, 2010). However, 'user satisfaction does not always correlate with high recommender accuracy' (Pu, Chen & Hu, 2011). In some cases, 'recommendations that are most accurate according to the standard metrics are sometimes not the recommendations that are most useful to users' (McNee, Riedl & Konstan, 2006). This divergence highlights how technical optimisation and user-perceived value can follow different trajectories, creating tensions in design priorities.

One of the most persistent challenges in RS development is the 'cold start' problem, often linked to data sparsity. These issues arise when there is insufficient information about new users, as 'to form any basis for recommendations' researchers rely on 'content information' (Schein et al., 2002). This reliance can result in 'recommendations for new items that appear similar to other recommended items,' which may limit the diversity of content presented to new users (p. 253). To address the cold start problem, researchers often draw on external data sources and apply knowledge transfer techniques (Pan & Yang, 2010). These approaches allow RS to model 'the way user and product

characteristics change over time, in order to distil longer term trends from noisy patterns' without requiring direct knowledge of the individual user (Koren, 2009). The next chapter shows that recommended content can also reflect the language associated with the IP circuit through which an account connects, understood here as the network routing path and inferred geographic location associated with an IP address at the time of access. This use of geolocation-linked language data is another method RS employ to address the cold start problem.

The use of personal data in RS raises privacy and security concerns. As users become more aware of potential risks and limitations in contemporary technologies, researchers and developers respond with increasingly sophisticated mitigation strategies. One early method to reduce the risk of users' 'loss of privacy' was to 'compute a public "aggregate" of their data that does not expose individual users' (Canny, 2002). In practice, this meant processing recommendations 'without exposing individual data' (p. 1). Current approaches build on these early aggregate-based techniques and include differential privacy, secure multiparty computation, and privacy-preserving data mining. These methods aim to preserve the quality of personalised recommendations while protecting the underlying user information. Yet, evidence from the Netflix Prize dataset shows that even anonymised recommendation data can be re-identified when combined with other publicly available sources, enabling researchers to trace records back to individuals (Narayanan & Shmatikov, 2008). This demonstrates that treating personal data in aggregated or de-identified form does not necessarily eliminate the risk of it remaining personal and traceable. These privacy considerations are not limited to data storage and transfer but

extend to the design choices within specific RS models. As later sections on collaborative and content-based systems show the type and scope of personal data collected condition how these systems generate and refine recommendations.

Privacy concerns are increasing as technological advances enable the use of more sophisticated forms of contextual data in RS. Beyond user and developer priorities, RS are influenced by the requirements of advertisers who fund platform operations. Contextual information such as time, location, and user interactions allows algorithms to adjust recommendations and create more adaptable environments for targeted advertising. For instance, ‘user-click feedback’ enables RS using a contextual bandit algorithm to adjust in order to ‘maximize total user clicks’ (Li et al., 2010). Li et al. detail how this approach can be applied to a news article recommendation service (p. 669). While RS can be designed to support user decision-making, the use of expanded contextual data in context-aware systems often serves to maximise user retention. In this role, the user becomes part of the attention economy, which ‘is toxic to important human values, because it harms individuals and society and it engenders and exploits weakened cognitive agency and vulnerability’ (Castro & Pham, 2020). This orientation towards engagement has also been associated with ‘negative effects on social utility’ (Milano, Taddeo & Floridi, 2020). In environments with large content databases, managing the balance between exploration and exploitation becomes a central element of the RS researcher’s decision-making process (Nguyen-Thanh et al., 2019). User-generated data such as ratings, clicks, and purchase history plays a central role in the collaborative filtering

approach to RS. Studies using this type of data have demonstrated improvements in algorithmic accuracy (Baltrunas & Amatriain, 2009). Baltrunas and Amatriain describe a 'micro-profiling approach' that applies a 'contextual pre-filtering technique' to recommendations, in which additional user data is analysed offline to refine the online processing performed by the RS (p. 5). In such cases, the initial configuration of the system may condition the range of content made available prior to the user's first interaction with the platform. While this method can enhance click-through rates, it also makes the capture, analysis, and modelling of offline user data a central and ongoing focus of RS development.

Collaborative filtering is among the most widely used approaches to building RS and can be implemented using either memory-based or model-based methods. The distinction lies in how each approach processes user interaction data. In a memory-based method, 'user-item ratings stored in the system are directly used to predict ratings for new items' (Desrosiers & Karypis, 2010). The system searches for similarities between users and items. If a user liked a particular item, that information can inform recommendations for other users. A user-based memory method can provide 'more accurate recommendations' but is less suitable in contexts where the 'number of users far exceeds the number of available items' (p. 140). In such cases, developers often adopt an item-based memory method. In Chapter 3, video-hosting platforms are analysed as cases in which the number of users exceeds the volume of available content, a condition under which item ratings play a prominent role in recommendation processes. Model-based approaches, in contrast, 'use these ratings to learn a predictive model' (p. 112). Here, a machine learning algorithm

processes the matrix of user interactions to predict recommendations based on predefined input parameters. These methods can yield more accurate predictions but require greater computational resources and longer processing times (p. 113). Both approaches have distinct affordances and constraints, and 'the "best" approach may differ from one recommendation setting to the next' (p. 140).

A content-based RS is another widely used model in RS research. Its core components are item features and user preferences. Unlike collaborative filtering, a content-based RS does not depend on feedback from a community of users, which can make the recommendations more directly personalised. One of its main limitations is that 'no content-based recommendation system can give good recommendations if the content does not contain enough information to distinguish items the user likes from items the user doesn't like' (Pazzani & Billsus, 2007). This means that if the user has not completed a detailed survey before creating an account, or if the survey questions do not align with the content's tags, the RS will not perform effectively. A content-based RS applies a learning algorithm to match a user's profile with item features, based on how those features are represented in the system's dataset (p. 339). This approach can be more transparent, as the user is generally aware of what information is provided during profile creation and how it feeds into the RS. However, user satisfaction with this method can be lower than expected, partly because it limits serendipity, which 'extends the concept of novelty by helping a user find an interesting item he might not have otherwise discovered' (Desrosiers & Karypis, 2010). In practice, recommendations often remain aligned with established

preferences, a tendency that can condition exposure to unfamiliar or unexpected items. Content-based RS do not face cold start or data sparsity problems. They operate by ‘matching up the attributes of a user profile, in which preferences and interests are stored, with the attributes of a content object (item) in order to recommend to the user new interesting items’ (Lops, De Gemmis & Semeraro, 2011). A key challenge with this model is the potential narrowing of exposure, which has been discussed in the literature as a condition associated with epistemic bubbles. Diversity in recommended content may exist in a general thematic sense, such as ‘public affairs’ or ‘baseball,’ yet this does not necessarily translate into diversity across ‘a democracy’s political spectrum of opinions’ (Haim, Graefe & Brosius, 2018: p. 2).

Both collaborative filtering and content-based RS rely on similarity and relevance measures to identify relationships between items. These relationships are quantified using distance techniques such as Euclidean, Jaccard, Cosine, Edit, and Hamming (Leskovec, Rajaraman & Ullman, 2014). These methods allow the RS to rank items and personalise recommendations, and are often applied to manage the balance between exploration and exploitation, a design trade-off that conditions patterns of content diversity within recommendation outputs. Content-based RS also track changes in the system using relevance feedback and active learning techniques. The rationale is the assumption that as ‘every user’s need is different and time varying, a database cannot adopt a fixed clustering structure’ (Zhou & Huang, 2003). Zhou and Huang observed that when ‘the size of the dataset is large and the dimensionality of the representation space’ on the platform is high, the ‘use of keywords or annotations (where

available) is key to meaningful retrieval' of information for improved recommendations (p. 542). By incorporating relevance feedback and active learning, RS can adapt over time to provide more accurate and personalised outputs that reflect the evolution of user interests. This adaptive process is discussed in the literature as a potential approach to mitigating certain forms of repetitive exposure, while not isolating or resolving the broader conditions associated with filter bubbles.

Researchers enhance similarity and relevance measures in content-based RS by applying learning-to-rank techniques. This approach directs computational power towards sorting 'new objects according to their degrees of relevance, preference, or importance' (Liu, 2009). It introduces an additional stage in the information retrieval process, which can improve efficiency. Learning-to-rank uses not only user–item interaction data but also contextual features of the item and demographic information about the user. As with the contextual pre-filtering technique discussed earlier, this additional data processing can limit the diversity of accessible content, even as it increases personalisation. Another way to refine similarity and relevance measures in content-based RS is by improving how user profiles and preferences are learned. In both cases, the overspecialisation problem can occur, where systems tend to 'find items that are similar to those the user has seen so far' (Aggarwal, 2016). To address this, researchers apply unsupervised learning methods such as clustering and dimensionality reduction. These techniques convert datasets 'to low-dimensional codes by training a multilayer neural network' (Hinton & Salakhutdinov, 2006: p. 504). Although this reduces the volume of data to

process, it can reveal relationships and patterns not apparent through conventional user–item comparison methods. This approach has been used to uncover ‘the real reasons of their [users’] rating behavior’ (Symeonidis, 2008). Although these techniques improve the precision and adaptability of content-based RS, they also illustrate how development decisions are shaped by trade-offs between personalisation, diversity, and computational efficiency. These trade-offs connect directly to the concerns addressed in the next section, where the evaluation of RS performance is examined in relation to commercial priorities and the allocation of resources during development.

Economy: Evaluation metrics, optimisation, and commercial priorities

Understanding the RS decision-making process can help researchers guide users towards personalised content. This level of precision is supported by quality measures, such as evaluation metrics, that assess the effectiveness of the technology. Quality measures are central to evaluation because users ‘do not accept high predictions if there is no reliability evidence’ (Bobadilla et al., 2018). In the digital economy, these measures become a priority for businesses, given ‘a direct relationship among e-service quality, e-satisfaction and e-loyalty in terms of online spending by consumers’ (Nisar & Prabhakar, 2017). Higher user satisfaction metrics can increase the profitability of a business developing RS, linking system quality to economic performance. The quality of RS is often measured using precision and recall metrics. For instance, these can identify cases where ‘certain users should not be included among the active user’s neighbours,’ which traditional evaluation metrics ‘are not capable of detecting’ (Ortega et al., 2013). RS rankings are based on different evaluation metrics, with

‘some measuring similar features, but some measuring drastically different quantities’ (Gunawardana & Shani, 2009). One method may measure ‘the distance between predicted preferences and true preferences over items,’ while another evaluates ‘the portion of favored items that were suggested’ (p. 2936). The choice of metrics is influenced by the context in which the RS operates and the strategic goals of the business. Because metrics guide how RS are tuned, prioritising one measure over another can shift recommendations to maximise commercial outcomes, making evaluation choices part of the system’s economic design. Economic factors also include development costs, market performance, and the competitive value of high evaluation scores, which can affect funding, shareholder confidence, and media visibility. The Economy section examines several evaluation metrics and their economic significance in RS development.

RS can be evaluated using four main categories of metrics: prediction accuracy, ranking, coverage, and diversity and novelty (Bobadilla et al., 2013). Each category provides a different perspective on RS performance. The metrics discussed here reflect a business incentive to expand catalogue coverage and diversity, even though such strategies can be less economically efficient than high-accuracy, convergence-oriented systems. In this context, the term refers to a RS optimised to minimise deviation from prior user behaviour. Such systems reinforce statistically dominant interaction patterns over time. While these metrics statistically capture user satisfaction, they cannot directly account for the quality or substantive relevance of the recommended content. Prediction accuracy metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Explained Variance Score (EVS) calculate the

difference between actual and predicted ratings for an item. These methods 'are frequently used for the evaluation of RS since they are easy to compute and understand' (Schröder, Thiele & Lehner, 2011), which makes them widely adopted and useful for maintaining competitiveness. When the primary goal is to maximise prediction accuracy, other considerations such as fairness, privacy, or transparency may be deprioritised in favour of improving quantitative benchmark performance. This can occur when 'most approaches [implemented by the researchers] only focus on the predictive accuracy' to improve performance metrics (Ge, Delgado-Battenfeld & Jannach, 2010).

Ranking metrics, including Precision, Recall, F1-score, Normalized Discounted Cumulative Gain (NDCG), and Mean Reciprocal Rank (MRR), assess how effectively a system orders items by relevance to the user. When these metrics are optimised alongside engagement signals, such as click-through rates, they have been associated in the literature with filter bubbles and echo chambers by reinforcing existing preferences and limiting exposure to diverse viewpoints, a dynamic described as a 'threat to democracy in the form of people's denial of opposing viewpoints' (Raza & Ding, 2022). While such ranking 'may be good for business, appears to serve users, and is relatively straightforward to implement,' it can also increase 'divisiveness, misinformation, out-group attacks, and addiction while drowning out critical information and deeper human connection' (Ovadya, 2022). Stray notes that research on ranking metrics often focuses on 'personalized user experience' without capturing the broader cultural effects of RS at scale (Stray, 2021).

Coverage, diversity, and novelty metrics, such as catalogue coverage, item-space coverage, intra-list diversity, and novelty score, evaluate the system's ability to provide varied recommendations. Coverage metrics can be divided into prediction coverage and catalogue coverage (Ge, Delgado-Battenfeld & Jannach, 2010). The refinement of 'prediction coverage measurement to take into account the usefulness of an item in a recommendation' (p. 258) aims to avoid scenarios where a system performs well statistically but does not deliver user benefit. A high score in such cases may indicate greater diversity within recommendation outputs, while unsuccessful suggestions can, through feedback loops in training data, be associated with the continued underrepresentation of certain items or users. Catalogue coverage faces similar limitations, as it measures how extensively the system uses the full set of available items, which does not necessarily improve recommendation quality. These limitations have prompted research into serendipity in RS design, where recommendations are both appealing and unexpected (p. 259). While this approach can enhance diversity, it often results in lower scores on other metrics, particularly for users with highly specific or narrow preferences.

Research on serendipity metrics remains limited in mainstream RS evaluation, with some studies suggesting that they may confuse users or reduce trust in the RS (p. 260). During the first decade and a half of RS development, academic and industry work largely prioritised accuracy metrics, though the emphasis varied across subfields (Castells, Hurley & Vargas, 2021). More recent attention to novelty and diversity metrics marks a shift from their earlier status as supplementary to relevance towards being treated as essential in business-

oriented evaluations (p. 605). This does not replace accuracy and relevance, which continue to dominate, but it broadens the range of performance measures considered valuable. Novelty and diversity are now presented as contributing directly to business objectives through product diversification, a recognised strategy for risk mitigation and market expansion (p. 606). Although novelty and diversity are ‘to some extent complementary dimensions’ (p. 604) and can support product diversification, they may also reinforce echo chambers when implemented without regard to content relevance. RS that optimise novelty or diversity using past interaction data may select items that differ in surface characteristics but still align closely with established preferences. This can preserve variation in form while maintaining the same underlying perspective, thereby reinforcing existing beliefs and excluding alternative viewpoints. Content producers, including news outlets and video channel owners, can influence novelty metrics by modifying titles or thumbnails to suggest unexpected or sensational content. This increases the likelihood that the RS will classify an item as novel, boosting its visibility in recommendations. These dynamics highlight that metric selection is not only a technical choice but also a strategic one, shaping how RS balance business objectives, user experience, and the diversity of recommended content.

Selecting appropriate metrics, or combinations of metrics, is a critical stage in RS development to support user engagement and satisfaction. Recent research shows ‘that 21% of all approaches do not include an evaluation and that 69% are evaluated using an offline evaluation’ (Zangerle & Bauer, 2022), drawing on large-scale literature surveys of peer-reviewed RS research papers,

particularly in the domain of experimental and applied RS models (Beel et al., 2013, 2016). The Framework for Evaluating RS proposed by Zangerle and Bauer contains ‘two main components: the evaluation objectives and the evaluation design space’ (p. 6). The objectives ‘are shaped by the perspective that is taken in terms of the recommender’s stakeholders’ (p. 7). RS evaluation typically involves multiple stakeholders, including platform owners, system developers, content providers, advertisers, end users, and regulators, ‘with varying goals and potentially conflicting interests, which may manifest in different evaluation objectives’ (p. 7). Stakeholders who create the recommended content ‘are a relatively new concern in RS research’ (p. 7; Gómez et al., 2022; Ferraro, Serra & Bauer, 2021; Ekstrand et al., 2018). The evaluation design consists of guidelines that ‘shape the method and setup of the evaluation’ and ‘should be considered and fixed early on in the process of evaluating a RS’ (Zangerle & Bauer, 2022: p. 8). This two-tier approach illustrates that ‘finding an adequate configuration for the comprehensive evaluation of a RS is a complex endeavor,’ particularly for platforms serving stakeholders with conflicting goals (p. 25). When these goals diverge, the choice of evaluation metrics determines whose objectives take precedence. Optimising for business metrics aligns RS with organisational economic priorities, often at the expense of objectives related to human development or public value (Tufekci, 2017). Metrics embedded early in development shape ranking logic, model optimisation, and the long-term behaviour of the system. This approach differs from multi-stakeholder or public-interest evaluation, as it defines success through measurable engagement or revenue indicators rather than broader societal outcomes. Its significance lies in

how it integrates economic priorities into the technical framework, creating forms of path dependence that can make later changes more resource intensive.

Loss Function

Examining how evaluation metrics are implemented during the post-training stage of the model shows the range of options available to software developers when selecting one recommendation strategy over another. These choices are often guided by RS objectives, such as to increase ‘the user satisfaction, increase an e-commerce website’s revenue, increase the number of items sold, sell more diverse items, help users understand the item space and engage users to increase their visit duration on a website, or return to the website’ (Zangerle & Bauer, 2022). All of the measures described above shape the optimisation criteria under which a trained model generates outputs. The algorithm that processes data in a trained model is shaped by the neural network (NN) during training through a loss function (LF). In many contexts, the LF is also referred to as a cost function, objective function, or error function (Goodfellow, Bengio & Courville, 2016), although the terminology and formulation may vary depending on the domain and application.

LF is a mathematical construct that allows developers to define many statistical problems ‘in the sense that loss functions define what a “good” estimator or a “good” prediction is’ (Hennig & Kutlukaya, 2007: p. 21; Lehmann & Casella, 2006). It ‘measures the performance of a learning algorithm’ (Beygelzimer, Langford & Zadrozny, 2008). Originating in statistics, the concept is often ‘described without a previous explanation of its functionality and

performance' in the literature (Bertolini & Schiozer, 2011). Within large statistical systems such as RS, the LF quantifies 'the "loss" in information that results' from the model's processing and estimation procedures (Ward Jr, 1963: p. 237). In the 1970s, 'the demonstration of the limitations of single-layer NNs was a significant factor in the decline of interest in neural networks' (Fausett, 2006). At that stage, efficient methods for developing NNs were lacking, and their potential applications had not yet been fully explored or realised. The later formulation of training algorithms 'played a major role in the re-emergence of NNs as a tool for solving a wide variety of problems' (p. 289). These algorithms incorporated LF, a practice that continues today. In stochastic systems such as RS outputs, the LF is used to assess how closely the system's predictions match the intended outcome, allowing the developer to compare the output of the current iteration with the target results.

Training a NN involves iteratively adjusting model parameters to reduce prediction error relative to a defined objective. This training is a complex task that can be very time-consuming for computers, because there is no simple formula for finding the best solution. A loss function acts as an evaluative measure, showing how far the network's current predictions are from the desired result. The aim is to adjust the network until this score is as low as possible. The training of a NN is classified as NP-hard, and specific formulations are NP-complete (Blum & Rivest, 1992). Richard Karp's work on computational complexity described such problems as those that can be verified efficiently but are not known to be solvable in polynomial time (Karp, 1972). When LF is defined within an NP-complete framework, solving it can involve computationally intensive

processes, often approached through brute force methods. In such cases, ‘the role played by the loss is marginal, and the choice of which loss to use for which problem is usually regarded as a computational issue’ (Rosasco et al., 2004). A long-standing ‘gap between the theory and practice’ in understanding the dynamics of deep NN learning began to close with the identification of nonlinear hyperbolic dynamics and the proposal of ‘exact solutions’ to address them (Saxe, McClelland & Ganguli, 2013). This work showed that relying on linear regression LF for NN optimisation was less efficient than assumed, as optimisation could stall at low-function values. Subsequent studies demonstrated that stagnation in learning was also linked to other factors (Dauphin et al., 2014), echoing earlier findings by Bray and Dean (2007). These stagnation issues indicated that the mechanisms underpinning NN training were not fully characterised, and earlier architectures lacked the transparency sometimes ascribed to them. In response, academic and industrial developers sought to accelerate training, for example by using stochastic gradient descent rather than processing all data at once, and by parallelising training on GPUs or TPUs to reduce the time to acceptable results. Recent advances in NN training algorithms ‘have the potential to directly influence both the production’ of systems and the way services are delivered (Cockburn, Henderson & Stern, 2018). Following the formulation of backpropagation and the resurgence of NN research, efforts concentrated on overcoming the obstacles that had slowed earlier progress. These conditions contributed to shifts in research focus and methodological priorities within the field. Because LF determine how models learn and what they optimise for, early design decisions can lock in constraints that shape later evaluation metrics, system performance, and the economic viability of the product. The discussion

of LF identifies gaps in the literature and traces their influence on data before accuracy metrics and performance standards are applied. By the time a researcher addresses challenges such as the exploration–exploitation trade-off, the NN model may already reflect parameters defined earlier in development and iteratively adjusted to align outputs with intended objectives.

LF is a fundamental component of the backpropagation (BP) algorithm. BP ‘is currently the most popular method for performing the supervised learning task’ and its main goal is to adapt an AI ‘so that actual outputs (P) come close to some target outputs (Y) for a training set which contains T patterns’ (Werbos, 1990). Originally introduced by Werbos in 1974, BP applies the principles underlying LF, although the function itself is not explicitly mentioned in the initial formulation (Hecht-Nielsen, 1992; Werbos, 1974). BP refers to a method of NN training that uses the LF, or cost function, as a reference value against which performance is evaluated. An NN ‘can be trained with backpropagation when desired output patterns exist and each function that has been used to calculate the actual output patterns is differentiable’ (Boden, 2002). Developers use the LF to compare these outputs with the target results and to evaluate the efficiency of the system. When the outputs of an RS show a strong linear correlation to the target values, the system is considered to be more effective for the chosen objective. Optimisation is achieved through ‘conventional gradient descent (or ascent), [where] backpropagation works by, for each modifiable weight, calculating the gradient of a cost (or error) function with respect to the weight and then adjusting it accordingly’ (p. 4). Although the use of LF in BP algorithms has proven effective in practice, research has paid more attention to its operational

benefits than to its broader implications. In commercially driven AI development, the practical advantages of using LF in BP tend to take precedence, while systematic evaluation of their potential wider impacts, including ethical considerations, remains limited.

The optimisation of an NN is shaped by the objectives of the RS, which address the needs of distinct user groups such as content creators, content seekers, and advertisers. The varying requirements of these groups, along with the system's output strategies, are balanced and adjusted during development by engineers using the LF in the training process. In the literature on RS optimisation, the LF is often presented as a practical means of measuring NN efficiency, embedding its value as a central motivation for intervening in and adjusting the AI's workflow at the programming level. This focus raises questions about how optimisation priorities are set and evaluated, particularly when broader impacts are not explicitly considered during system design. For instance, in Konrad Kording's work on building an NN 'to describe unsupervised learning of neurons in brains,' the LF created an ethical dilemma because its dependence on input data and neuron parameters meant that optimisation outcomes could be influenced by experimental design choices rather than reflecting biological reality (Kaznatcheev & Kording, 2022). Kording recognised that if the LF reinforced patterns introduced by the experimental setup rather than those genuinely present in neural activity, it could lead to misleading conclusions about brain function, potentially distorting scientific understanding and guiding future research on a false premise. To address this, Kording developed a separate system to differentiate 'the loss to calculate the gradients' (p. 7). Such thorough

approaches are necessary not only in biological engineering and neuroscience but also when processing information where low quality or bias can condition the informational environments relevant to civic participation. This example highlights the analytical importance of examining the origins of such algorithms and the implications of their application.

Recent research identifies bias in RS as being linked to practices that ‘over-sample the “difficult” instances in order to increase learning speed,’ particularly when LF is used to support this process (Chen et al., 2020). Modern LF designs for deep face recognition, for example, are often evaluated only in terms of ‘training accuracy, test accuracy, and convergence rate’ (Srivastava, Murali & Dubey, 2019). The decisions by companies such as Meta and Microsoft to halt certain facial recognition projects in 2021 reflected concerns about ethical risks and potential misuse, especially in surveillance and law enforcement (Milmo, 2021; Bloomberg, 2020). In such cases, the risk was not inherent to the algorithm itself but arose from the nature of the data and its possible applications. LF is described as ‘the most widely publicized feature in the devices today and hence represents an important problem that should be studied with the utmost priority’ (Srivastava, Murali & Dubey, 2019). However, Srivastava, Murali & Dubey, along with other authors cited in this chapter, do not examine the implications of data that may be created, transformed, or excluded during training when LF is used to guide optimisation. For instance, Heng-Tze Cheng et al. address the problem of RS recommending ‘less relevant items’ (Cheng et al., 2016) by proposing a model ‘trained to minimize the final loss function during model training’ (p. 2). This approach made the RS more efficient and allowed LF

to be used as a tool for comparing recommendation outputs, yet it did not consider the broader consequences of applying LF to training data that had been shaped or filtered during model development. The experiment led to a ‘significant improvement on app acquisitions’ on Google Play (p. 3), meaning that more users installed the promoted applications. While this demonstrated the commercial effectiveness of the model and increased the business value of the RS, it left unexamined potential impacts on information diversity, user autonomy, and exposure to biased content.

The studies reviewed in this chapter each use a different variation of the LF, defined by the authors for a specific task. In each case, the LF is adapted to improve computational performance, yet its broader implications for the system are not examined. For instance, Zhang et al. provide a ‘taxonomy of deep learning-based recommendation models’ and a ‘comprehensive summary of the state of the art’ (Zhang et al., 2019), noting that in this field ‘authors generally have free rein over the choices of datasets/baselines’ and proposing a ‘suite of standardized evaluation datasets’ (p. 29). Such an approach could enable more systematic comparison of practices and support closer examination of how changes in RS versions relate to differences in data handling. However, the diversity of programming challenges often necessitates unique datasets for each task. Zhang et al. also raise questions about the ‘difficulty of test samples in RS results’ and whether the ‘chase-after-time’ approach is appropriate (p. 29). Despite this critical framing, their analysis of several LF compares them only in terms of computational efficiency. Similarly, Ying et al. apply a Graph Convolutional Network to ‘make high-quality recommendations’ and describe the

proposed ‘margin-based loss function in detail’ (Ying et al., 2018). While they explain the formulation of the LF, they do not explore its possible effects on the data being processed. Across the literature reviewed here, in-depth discussion of the ethical risks associated with brute force training processes in NNs remains limited. Although the ethical implications of LF use are broadly acknowledged in abstract terms within the software development community, they remain largely unaddressed in practical implementation.

The concept of LF is important for examining how economic priorities are translated into technical optimisation choices during RS creation and implementation. The way in which a developer defines and formulates the function determines how the original code is adjusted during the training stage of the software, with performance assessed against the values generated by the LF for that specific system. However, the approach selected in each case often does not explicitly consider the potential effects of optimisation on the training data that may be overlooked or excluded. For this reason, it is necessary to clarify the values and processes that underpin optimisation using LF, and to consider the ways in which these values become embedded in the RS produced through such methods.

The LF used by developers can be broadly classified into two categories: regression and classification (Muthukumar et al., 2021). A regression LF is applied when predicting continuous values, such as the distance between a predicted value and the actual value for a given recommendation. A classification LF is used when predicting categorical outcomes, such as class labels or binary

decisions. The regression approach enables developers to assess how closely the RS output matches the predicted item for a specific user based on the dataset. The classification approach narrows recommendations to groups of items under a shared category, such as a type of product or style of music, and incorporates binary decisions such as positive or negative ratings, or purchased and not purchased items. Considering the differences between these two types of LF is essential for selecting the training process that aligns with the RS's objectives. This choice influences how the NN's training is evaluated and embeds a particular pattern of data processing into the software (Barinov et al., 2023). The decision between regression and classification LFs shapes how data is processed and represented within the system.

Several regression LF are commonly used in RS, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and Huber loss. MSE is widely applied when prediction values take the form of user–item ratings. Both MSE and MAE were mentioned earlier in this chapter in the context of prediction accuracy metrics, where they were used to evaluate how well a trained model processes unseen datasets. In such cases, these same measures function as LF. While 'squared loss [MSE] is a standard choice for the loss function,' MAE can be more suitable when error magnitude is linear, interpretability is important, and outliers are frequent (Melville & Sindhwani, 2010). MSE squares all errors, giving more weight to larger deviations, whereas MAE treats all errors equally. When large errors in recommendations are more costly than smaller ones, MSE is often preferred. In e-commerce, where much RS development and research takes place due to the availability of large datasets and extensive user bases, the high

cost of errors for the business often leads to a preference for MSE (Cui, Wong & Wan, 2012). This focus on minimising large errors can influence coverage, novelty, and diversity, depending on the type of content recommended. For instance, an online retailer might prioritise accurate recommendations for high-value electronics over experimental suggestions, as mistakes in such cases could result in substantial revenue loss. Emphasising MSE can also lead to popular items receiving disproportionately high importance scores, since errors on these items may carry greater business costs, such as losing a customer. A further limitation of MSE is its sensitivity to outliers, which have a strong influence on results. Reducing their effect can limit how extreme or atypical values are represented within the model, conditioning the range of information available for interpretation. For instance, in news recommendation, deprioritising articles with perspectives outside the dominant narrative may reduce prediction errors while narrowing the range of viewpoints surfaced within recommendation outputs, shaping the informational context in which political communication occurs. The implications of these optimisation choices are examined across Chapters 3 to 6, tracing how politically distinct search phrases on video-hosting platforms shape recommendation patterns, how these align with AI policy frameworks, how they reflect technical design choices, and how optimisation interacts with governance, regulation, and socio-economic contexts, shaping information visibility and epistemic agency.

The third regression LF is Huber loss, which combines MSE and MAE by switching between them when the error exceeds a defined threshold. In this way, Huber LF as a compromise between the two methods, automating the transition

between them. However, it does not resolve the limitations of MAE with respect to large errors and is non-differentiable at the switching point (Gupta, Hazarika & Berlin, 2020). Huber loss is also more computationally complex and requires longer training times, which can make it less viable in economic terms. Its application of multiple conditional rules makes it harder to interpret than MSE or MAE (Leski, 2014). The choice of the switching point introduces an outlier trade-off: setting it too low can reduce coverage, diversity, and novelty, whereas setting it too high increases the influence of regular errors on recommendations and can reduce user satisfaction. For example, a video-hosting platform could set the Huber loss switching point too low, treating a broader range of atypical items as outliers, thereby narrowing coverage and novelty within recommendation outputs. If set too high, the system focuses on small errors within familiar topics, amplifying already dominant narratives and reducing exposure to alternative perspectives. In both cases, the trade-off affects how much epistemic range users experience in their recommendations. These considerations for regression LF highlight how methodological choices in optimisation shape both system behaviour and user experience. The same is true for classification LFs, which apply a different set of principles to categorise recommendations and guide RS outputs.

Classification LFs such as Hinge loss, Cross-entropy loss, and Kullback–Leibler (KL) divergence are widely used in RS research. Each presents potential challenges that can contribute to the creation of an epistemic agency gap, giving added weight to the developer’s choices during the AI training stage. Hinge loss is used for both multi-class classification and binary tasks, such as in Support

Vector Machines, and measures the distance between the true and predicted labels. Although ‘for classification, the hinge loss appears to be the loss of choice’ (Rosasco et al., 2004) and it is widely applied in practice (Molitor, Needell & Ward, 2021), the method focuses on data points lying close to the boundaries between classes. This emphasis means that data considered well-classified is largely ignored, limiting its influence on the RS’s decision-making process. Such omissions can condition the visibility of content associated with underrepresented groups within recommendation outputs, shaping the informational context in which public communication takes place. Mitigation strategies for such effects are discussed later in the Culture section of this chapter. Cross-entropy loss, also known as logarithmic loss, is ‘popular mainly due to its excellent convergence speeds, alongside its excellent performance’ (Ho & Wookey, 2019). It measures the dissimilarity between the actual and predicted distributions of recommendations, which can lead to underrepresented classes, those with small distributions, being marginalised further in model outputs. Moreover, when the LF makes a confident but incorrect prediction for a rare class, its logarithmic form amplifies the influence of this error, potentially creating an epistemic agency gap by limiting access to certain content. These effects occur when class imbalance and decision thresholds are not managed. KL divergence measures the difference between a predicted probability distribution and a reference distribution. A limitation of this LF is that it ‘treats all class probabilities in the same way,’ merging non-dominant hypotheses into a single event (Ponti et al., 2017). This averaging process can obscure meaningful distinctions between classes, diminishing the weight of knowledge that may be relevant for well-informed decision-making.

The LFs discussed for regression and classification are standard industry methods that can introduce some degree of information misrepresentation from the dataset, either through the mathematical approach used or because the dataset is not adapted to the chosen technique. Analysing the developer's decision-making process when selecting an LF can reveal the economic considerations behind this choice, as linking model behaviour and objectives to economic outcomes shows how the direct and indirect costs of RS errors influence which LF is applied. Software companies that employ RS derive part of their value from their 'abilities to implement price discrimination,' which can drive the creation of 'firm-specific data sets and applications' of AI (Agrawal, Gans & Goldfarb, 2019). This highlights the need to recognise that 'the data we [researchers] are analyzing is going to be the result of decisions that are made by algorithms in which the decision-making process may or may not resemble the decision-making processes we model as social scientists' (p. 28). The choice of an LF during RS model training is a design decision through which economic priorities can be translated into technical optimisation criteria. Developers select an LF not only for its statistical properties but also for its capacity to produce outputs aligned with company objectives, such as higher engagement, increased sales, or greater advertising revenue. This alignment is rarely neutral. When a platform seeks to maximise a financially significant metric, such as click-through rate or predicted willingness-to-pay, the LF is tuned to prioritise patterns in the data that generate those results. The model then learns to emphasise signals that are profitable to the business while de-emphasising those that do not contribute to the targeted metric. This process influences which content is

surfaced, ranked, or suppressed, shaping the overall circulation of information in the system. The result is a set of positive and negative biases, where certain types of information are amplified while others are marginalised. Over time, such alignments can shape user experience by prioritising content consistent with commercial objectives, potentially conditioning the range of epistemic variation available within the system. For instance, LFs that disproportionately penalise large prediction errors, such as MSE, can overfit to popular items that produce bigger errors when mispredicted, increasing their ranking frequency and reducing exposure for less common or niche items. In this way, the LF embeds a commercial logic into the architecture of the RS, favouring content with stable, high-volume engagement over content with sporadic or low-volume interaction. The gap that emerges is not simply the absence of certain information, but a systematic omission shaped by the alignment between LF design and economic incentive. This gap becomes harder to address over time because the LF's influence is cumulative: once trained, the model reflects its optimisation history, and retraining with a different LF requires re-engineering datasets, pipelines, and evaluation frameworks

Businesses often design RS to support price discrimination, aiming to 'quickly develop the capacity to personalize prices' by targeting users who 'do know a lot about what they want and have precise willingness-to-pay' (p. 604). While this approach meets consumer goals by delivering desired products at optimised prices, it can conflict with the user's role as a citizen by reducing interaction to transactional value and excluding collective interests, public accountability, and epistemic diversity. Economic pressures to release features

quickly add to technical debt (TD), where immediate deployment takes precedence over mitigating representational bias. TD, as described by Cunningham (1993 [1992]), refers to the long-term costs of maintaining and updating software due to factors such as poor code quality, incomplete testing, or inadequate documentation. Although TD 'is recognised as a critical issue in the software development industry,' the 'absence of a clear definition and model for technical debt exacerbates the challenge of its identification and adequate management' (Tom, Aurum & Vidgen, 2013). Addressing TD 'is not to add new functionality, but to enable future improvements, reduce errors, and improve maintainability' (Sculley et al., 2015), which reduces operational costs such as developer time, infrastructure load, and system failure management. The choice of LF and its potential effects on information representation are rarely included in TD considerations. By the time TD is addressed, the model is typically already trained and deployed, making such information gaps difficult to reverse in practice without substantial reconfiguration of data pipelines, evaluation frameworks, and organisational priorities. Although retraining costs have declined, recent work shows that retraining decisions are governed by explicit trade-offs between model staleness, resource cost, and downstream performance, making retrospective correction of representational choices difficult once systems are deployed (Mahadevan & Mathioudakis, 2024). Selecting an LF that mitigates risks from sensitive information requires additional tuning, testing, and infrastructure, which increases development time and cost. Industry reports often frame this as a trade-off between speed to market and model robustness, with rapid deployment prioritised over incremental improvements in coverage or bias reduction.

There are several trends in the research literature that aim to address limitations imposed by the LF. Robust optimisation, fairness-aware LFs, and meta-learning or learning to optimise are approaches that focus on improving model performance during training and convergence. Robust optimisation adjusts the LF to account for the worst-case training outcome, reducing the influence of data points far from predicted values. It is ‘a specific and relatively novel methodology for handling optimisation problems with uncertain data’ (Ben-Tal, El Ghaoui & Nemirovski, 2009), where uncertainty refers to ambiguities in interpretation or inherent randomness, such as those found in human behaviour data. This method can over-emphasise noise and outliers, which may reduce performance on average cases and has been discussed as a potential source of bias amplification under certain data conditions. Fairness-aware LFs reflect efforts to reduce discriminatory patterns in model training. While they can be applied to deployed models without altering them, they are often impractical in contexts such as commercial APIs, where access to deployed models and their intermediate outputs is restricted (Du et al., 2021). Meta-learning and learning to optimise is ‘an emerging field that aims to optimize various parts of learning by infusing prior knowledge of a given task distribution such that new tasks drawn from the distribution can be solved faster and more accurately’ (Psaros, Kawaguchi & Karniadakis, 2022). This approach uses parameters derived from prior data to accelerate adaptation, which also allows biases and uncertainties from earlier iterations to influence subsequent training. Together, these approaches demonstrate how technical adjustments seek to mitigate known constraints, while also introducing new dependencies and trade-offs that shape

the next stage of RS model development. These technical adjustments operate within broader cultural frames that influence how optimisation is interpreted and prioritised, a relationship examined in this thesis when analysing how such methods are applied.

Culture: Data provenance, diversity, and ethical risks

As discussed in the Economy section, the selection of evaluation metrics and LF in AI training is often shaped by uncertain, incomplete, or biased data, which can reduce accuracy, skew recommendations, or limit coverage. Concepts such as data bias and data provenance have therefore become central to RS research and practice. In this chapter, culture is examined at the level of data practices and infrastructural conditions that shape how information is collected, processed, and rendered visible within RS, rather than at the level of media content, political discourse, or creative production, which correspond to Webster's cultural and political dimensions at the level of representation rather than infrastructure. These latter dimensions of culture, including the role of RS in news circulation, political communication, and worldview formation, are addressed in later chapters through empirical case studies and policy analysis. The focus here is on upstream design choices that precede content exposure and condition the range of cultural material that can enter recommendation pipelines. This section therefore examines current developments in data bias and data provenance, focusing on how these factors influence AI systems before training and evaluation, and how such upstream choices affect the information made visible and accessible to users.

Data bias and data provenance are central factors affecting the accuracy, transparency, and efficiency of RS. Although many examples originate from domains outside RS, the underlying mechanisms are comparable, as RS rely on similar processes of data collection, processing, and inference. Bias in machine learning is defined as ‘inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair’ (Ntoutsi et al., 2020). Because algorithms process data through mathematical and statistical abstractions, they can prioritise certain classes over others, making bias a key concern in ethical AI discourse. Data provenance is described as ‘where a piece of data came from and the process by which it arrived in the database’ and is increasingly recognised as critical, particularly in scientific contexts where accuracy and timeliness are essential (Buneman, Khanna & Wang-Chiew, 2001). Data provenance can be examined from three perspectives: agent-centred, which focuses on ‘what people or organizations were involved in generating or manipulating a resource’; object-centred, which focuses on ‘tracing the origins of portions of an entity, i.e., an object or a resource, to other entities;’ and process-centred, which focuses on ‘capturing the activities and steps taken to generate a resource’ (Batini & Scannapieco, 2018, p. 429). In RS, documenting anonymisation processes forms an important part of provenance practice, linking privacy protection with the ability to verify the origin and handling of data. As a form of metadata, provenance records both the source and the processes that produced the data, and is ‘also referred to as lineage and pedigree’ (Simmhan, Plale & Gannon, 2005). Both bias and provenance shape RS development and condition the contexts in which social and ethical implications of deployed systems are discussed.

In the context of RS, data bias and provenance shape recommendation relevance and the informational scope within which users engage with system outputs. Policy makers and regulators often prioritise features such as explainability and transparency to improve how users interpret AI outputs. For instance, a RS may rely on ‘behavior data [which] is observational rather than experimental’ and in which ‘biases widely exist ..., including but not limited to selection bias, position bias, exposure bias, and popularity bias’ (Chen et al., 2023). Biases can arise from multiple sources, including behavioural data, uneven presentation of items, or feedback loops where user responses are used to refine recommendations. In the latter case, the process can introduce new biases and amplify existing ones. Despite the ‘wide usage of the terminology “bias” in the literature, its definition is usually vague and even inconsistent across papers’ (Chen et al., 2021). As a result, work on explainability and transparency often extends to accuracy, robustness, and other aspects of system behaviour, with data bias and provenance forming part of this broader category. Such research aims to improve the conditions under which recommendations are generated while acknowledging the trade-offs that accompany attempts to make them fairer and more reliable.

This section examines data bias and provenance within their cultural context and their influence on RS outputs. Cultural context shapes how these factors condition information environments relevant to civic participation. As noted earlier in this chapter, challenges linked to bias and provenance emerge at the starting point of RS development, with their effects carried through the

value chain to the user experience. Seven types of bias are considered: historic, sampling, labelling, proxy, data imbalance, contextual, and lack of diversity. The discussion addresses data collection and acquisition practices, with attention to preprocessing and cleaning. These elements form core components of the Extract, Transform, Load (ETL) framework that underpins AI training. Key concepts, challenges, and potential solutions relating to bias and provenance in RS are explored, alongside consideration of evaluation metric design described earlier in the chapter.

In 2023, with the introduction of large language model services such as ChatGPT, based on OpenAI's GPT-3.5 and GPT-4, and Gemini [Bard], based on Google's LaMDA, media reports and user discussions raised concerns about biased or misleading outputs linked to the data used for training (Wiggers, 2023; Grant & Weise, 2023). It also became public that these datasets included unlicensed materials, prompting legal action from artists, stock image providers such as Getty Images, and writers including Sarah Silverman, who sued OpenAI and Meta AI for alleged copyright infringement during training (Brittain & Brittain, 2023; Vincent, 2023; Milmo, 2023). Companies often respond to bias and provenance concerns by applying filtering layers to already trained models, a faster and less resource-intensive alternative to full retraining. This approach can affect model accuracy, however, underscoring how mitigation strategies interact with performance. Modifications to GPT-4 between its March and June 2023 releases illustrate this trade-off: while the updated version of ChatGPT 'out-scores medical students on complex clinical care exam questions' (Hadhazy, 2023), its accuracy in identifying prime numbers fell from 97.6% to 2.4% (Chen,

Zaharia & Zou, 2023). Such changes indicate that updates intended to address bias or other issues can reduce accuracy in unrelated domains. Identifying these issues early in development highlights how interdependencies between stages can complicate later mitigation efforts.

The first type of bias considered here is historical bias. In this context, training datasets are shaped by past policies, societal norms, and cultural practices embedded in both the data and the processes used to collect it. These influences can distort representations of current events and affect the predictive capacity of RS. For instance, an AI trained on prior hiring data from an employer that systematically rejected candidates from women's or historically black colleges 'may learn to discriminate female or black applicants' (Barocas & Selbst, 2016). Historical bias can also reflect cultural differences without involving socially or politically sensitive information. If 'people from some regions may simply not have purchased the product because it wasn't previously marketed in their region,' a RS may reinforce this gap by continuing to classify the item as unpopular (Roselli, Matthews & Talagala, 2019). When datasets reflect historical trends and RS are optimised to maximise engagement by recommending items with high past popularity, recommendation outputs may align more closely with previously dominant choices and behaviours. This can limit exposure to less represented items, perspectives, or communities. In collaborative filtering approaches, discussed earlier in this chapter, past user behaviour is used by design, making systems particularly sensitive to historical bias in the training data. As noted by Barocas and Selbst, datasets containing historical bias 'can deny historically disadvantaged and vulnerable groups full participation in

society’ and, when discrimination stems from intentional past exclusion, ‘there is frequently no obvious method to adjust historical data to rid it of this taint’ (Barocas & Selbst, 2016). The case study of the Russian video-hosting platform in the next chapter illustrates how politically distinct search terms were associated with narrower recommendation sets for some viewpoints, highlighting patterns of differential visibility within RS outputs. Addressing historical bias requires careful data auditing and considered selection of the LF used in training. As Barocas and Selbst argue, such datasets ‘force society to explicitly rebalance the two justifications [complications internal to the logic; external political and constitutional difficulties] for antidiscrimination law’ and ‘make it simply impossible to rectify discriminatory results without engaging with the question of what level of substantive inequality is proper or acceptable in a given context’ (p. 676). These challenges are complex and persist across domains, whether arising from explicit past discrimination, geographic gaps in product marketing, or entrenched majority preferences in fast-moving, commercially driven technological markets.

The second type of bias considered here is sampling bias. This occurs when samples selected to represent a dataset do not proportionally reflect all relevant categories, such as user demographics, item types, or decision outcomes. It ‘arises due to non-random sampling of subgroups’ (Mehrabi et al., 2021) and can result in RS outputs where ‘the trends estimated for one population may not generalize to data collected from a new population’ (p. 7). Sampling bias can originate during the selection process or from non-responsive samples. An example of the former appears in clinical studies, where

convenience sampling from urban sites over-represents street-active transgender women, such as those accessing HIV testing services (Bauer & Scheim, 2013). Similarly, 'convenience samples' have contributed to 'the demographic imbalance of popular public face datasets in terms of identity, gender and ethnicity,' which has been linked to bias in face recognition systems (Hupont & Fernández, 2019). Sampling bias often emerges when processes aim to generalise from the training dataset as a whole. In RS development, sampling bias can also emerge when training data disproportionately reflects high-popularity items, a condition under which commonly used evaluation metrics and LF may align more closely with dominant subsets of the data rather than the dataset as a whole.

Labelling bias arises when human judgement influences the data collection process. In RS datasets, items may be labelled through explicit user ratings or implicit signals such as behaviour. Mislabelling can result from cultural or subjective interpretations, leading to skewed recommendations. For instance, OpenAI's outsourcing partner Sama 'employs workers in Kenya, Uganda and India to label data for Silicon Valley clients like Google, Meta and Microsoft' (Perrigo, 2023). When labelling is conducted without a consistent codebook, cultural context can shape classification decisions. Such labels may determine which items are identified as toxic and influence how bias mitigation measures are applied, thereby constraining AI outputs. Outsourcing labelling to regions with different cultural contexts can embed both cultural and sampling bias, as annotators apply local norms to classification tasks. This affects how categories such as toxicity or relevance are operationalised within the system, conditioning

which perspectives are more or less likely to be represented in model outputs. Labels may also carry different significance for users than for the RS. A user might see certain attributes as central to their identity or interests, while others are trivial or situational. For instance, “dog owner” may be recognisable as significant to a user, while “bought a novelty sweater” would be less socially significant; yet the RS may still regard it as statistically significant when making inferences about the preferences of the user’ (Milano, Taddeo & Floridi, 2020). In such cases, the RS evaluates attributes solely on statistical association, without accounting for their social or personal relevance, which can result in recommendations misaligned with user priorities. Approaches such as assembling diverse annotator pools and adopting consensus-based labelling can reduce the effects of labelling bias, although these require substantial economic investment from developers.

Proxy bias arises when indirect correlations between sensitive and specific data influence model outputs. This can occur when attributes such as postal codes are strongly associated with particular demographic groups, leading to discriminatory outcomes. For example, the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software was criticised for using postal codes as a proxy for the likelihood of crime or recidivism. Because postal codes often correlate with neighbourhood demographics, including racial composition, their combination with past arrest records produced a bias that disproportionately classified black defendants as higher risk (Angwin et al., 2016). This approach ‘was biased against black defendants as the group was associated with a higher false positive rate’ due to the model’s use of indirectly

related variables as proxies for one another (Srinivasan & Chander, 2021). Such proxies are learned during training when statistical correlations are operationalised within the model, reinforcing spurious patterns in the data. Even with bias and provenance checks, models designed for efficiency may adopt correlations as shortcuts, which can act as proxies for sensitive attributes. Mitigating proxy bias requires identifying and testing for indirect correlations, including those not overtly linked to sensitive variables. In large-scale systems, this often involves manual inspection or human-assisted labelling to determine whether variables serve as unintended proxies. This process introduces its own challenges, as annotators may apply subjective interpretations, organisational practices, or cultural assumptions that embed further bias. In practice, the choice is between the opacity of automated detection and the subjectivity of human judgement, neither of which guarantees the removal of proxy effects.

Data imbalance occurs when classes are unevenly distributed within a dataset. Like sampling bias, it can result in the underrepresentation of minorities, raising questions about equality and equity in RS outputs. Research has addressed major and minor imbalances through n-fold selection and large-scale imbalance reviews, yet these approaches adjust class distributions without resolving the structural conditions that generate imbalance. As a result, bias can re-emerge when models are applied to new or real-world data. This bias is relevant to the case studies of video-hosting platforms discussed later, where uneven representation across content categories and user attributes influences both the visibility of material and the diversity of recommendations. It also links to earlier discussions of how RS evaluation metrics and LF interact with sparse

or skewed datasets, and to later chapters on how structural biases in training data shape epistemic diversity. On social platforms, sparsity describes the high proportion of missing or unrecorded values in user–item interaction matrices, which can occur when users leave profile fields blank, interact with only a small portion of available content, or maintain few network connections (Pan, He & Yu, 2020). Such imbalances can lead a model to output zero for underrepresented classes, reducing recommendation quality. For instance, in a binary mammography dataset, the overrepresentation of negative cases compared to positive ones produced an imbalance bias in which the majority class achieved close to 100% accuracy while the minority class reached only 0–10%, resulting in approximately 234 misdiagnoses (He & Garcia, 2009). Because social data is often imbalanced, some groups or categories remain underrepresented, and when evaluation metrics and LF are applied to such data, this imbalance can condition model reliability and the representational scope of system outputs. Mitigation strategies often seek to flatten class distribution, but, as in the case of historical bias, doing so raises broader questions about ‘what level of substantive inequality is proper or acceptable in a given context’ (Barocas & Selbst, 2016).

Contextual bias arises when a dataset reflects the specific circumstances of its collection, limiting its ability to capture the full range of diversity and variation. This bias embeds the characteristics, constraints, and dominant patterns of the collection setting into the dataset, reducing the adaptability of models trained on it. Such models may underperform when applied to contexts with different cultural norms, temporal patterns, or user characteristics, for example when a system trained on urban daytime music preferences is deployed

in rural night-time listening contexts or for demographic groups absent from the original data. Contextual bias can contribute to feedback loops in which misinterpretations of certain items restrict content exposure, conditioning the informational range available within recommendation outputs. In music recommendation, for instance, the relevance of results may differ between day and night, affecting user satisfaction. Whether 'an algorithm can be considered discriminatory or not depends on the context in which it is being deployed and the task it is intended to perform' (Ferrer et al., 2021). Cases 'in which an algorithm is biased toward hiring young people...it can be considered that the algorithm is discriminating against older people' (p. 73) illustrate how bias is shaped by context, task, and perspective, as the same output can be interpreted differently depending on whose position is considered. If the context in which the model is deployed does not justify hiring more young people than older people, it may be recognised as contextual bias (Hao, 2019). This type of bias 'is an under-addressed kind of bias in the legal AI literature' (Price & Nicholson, 2019). While 'contextual bias needs addressing, policymakers should not be misled by the Nirvana fallacy' (p. 69), where solutions are dismissed for being imperfect, even partial mitigation can bring societal benefits, such as expanding exposure to underrepresented content, improving relevance for marginalised groups, or reducing reinforcement of narrow recommendation loops (p. 70).

Lack of diversity is a distinct form of bias. Unlike sampling bias, which arises during data collection, or data imbalance, which emerges during model training, lack of diversity is identified during dataset analysis and reflects the underrepresentation of specific demographic groups. Addressing it can

significantly improve the quality of AI outputs. This bias is particularly relevant in the medical domain, where insufficient reporting on patient diversity can hinder the clinical application of AI (Daneshjou et al., 2021). For instance, datasets from the International Skin Imaging Collaboration lack skin tone information, although external groups have attempted to enhance labelling to capture greater diversity (Kinyanjui et al., 2019). Without such information, developers cannot make informed decisions about whether and how to address diversity gaps to improve model performance. In the medical context, ‘assessing racial and ethnic diversity among patients in these data sets was limited by sparse reporting of these characteristics,’ a limitation that increases the risk of inaccurate recommendations with potentially significant consequences for patient outcomes (p. 1366). The influence of lack of diversity bias is also relevant to civic participation on video hosting platforms. ‘Preferences and opinions differ between groups of citizens with lower and higher levels of educational attainment,’ shaping ‘the topics that are on the table during mini-publics and the proposals that they produce’ (Binnema & Michels, 2022). In environments where ‘focus on language, deliberation and arguments potentially enhances the dominance of the higher educated,’ video hosting platforms ‘do not compensate for this lack of diversity among the participants’ and this ‘has an impact on the issues and preferences put forward – and their uptake by politicians and civil servants’ (p. 478). Mitigation strategies can include promoting inclusivity during data collection, enforcing diversity metrics, adapting LF during training, and modifying trained models to enhance output diversity. These measures are discussed in the literature as approaches that may shape information

environments relevant to civic participation, while remaining subject to commercial and performance constraints.

Data sources and their diversity influence the transparency and accountability of a dataset. In RS, user feedback and behavioural data are the primary sources, and these can be categorised as ‘social knowledge about the user base in general,’ ‘individual knowledge about the particular user for whom recommendations are sought,’ and ‘content knowledge about the items being recommended’ (Burke, Felfernig & Göker, 2011). Diversity across these knowledge types can support the production of recommendations that are more nuanced and contextually differentiated. However, the same diversity that enhances RS outputs can also complicate the data collection process. Data gathering often relies on ‘trust and homophily [which] both play an important role in the social recommendation process’ and ‘in some cases, trust is the consequence of homophily’ (Aggarwal, 2016). On a video hosting platform where links may reflect interests and political views, both factors are relevant, yet excessive homophily can condition patterns of exposure by aligning recommendations more closely with existing preferences. Many RS, ‘in both academic and commercial settings, obtain ratings by having users explicitly state their preferences for items’ as this is ‘a straightforward method and avoids potentially difficult inference problems for user preferences’ (Ekstrand, Riedl & Konstan, 2011). Managing large volumes of information, addressing sparse or noisy data, protecting user privacy and security, mitigating the cold start problem, and adapting to the temporal dynamics of user preferences all shape how these systems collect and use information. Each of these factors involves trade-offs

between the richness of collected data and the degree of intrusion into user activity, making a systematic approach to collection strategies essential.

Developers employ various techniques to strengthen data acquisition, including UX designs that encourage engagement and feedback, a mix of implicit and explicit collection methods, gathering diverse data to capture multiple aspects of user behaviour and preferences, and integrating privacy and ethical safeguards. Data acquisition presents a high entry barrier, 'which makes it hard for new service providers to compete with existing recommendation services in a domain' (Heitmann & Hayes, 2010). One approach to reducing this barrier is the use of linked data 'to mitigate the high cost of data acquisition in closed RS,' enabling developers to 'fill in the gaps' in the background data of the RS (p. 79). Linked data is a framework for connecting related information across the internet, typically implemented through the Resource Description Framework (RDF) and other semantic web standards that promote common formats and exchange protocols. While this can reduce the time and computational resources required for data acquisition, it can also introduce previously discussed biases into the dataset, for example by aligning linked sources with similar characteristics.

Data pre-processing and cleaning are essential for converting raw data into a refined dataset suitable for use in a RS. These processes support the quality and reliability of recommendations, as measured by predefined evaluation metrics. Data cleaning involves detecting and correcting errors such as inconsistencies, duplicates, and inaccuracies, typically in two stages: error detection and error repair. This is important because 'data quality remains to be

a major concern, and dirty data can lead to incorrect decisions and unreliable analysis' (Chu et al., 2016). In some cases, cleaning can alter or obscure genuine relationships between items, making it a resource-intensive process that incorporates human judgement. Data integration merges information from different sources, but discrepancies in collection methods can introduce inconsistencies that, in turn, create bias and lead to misleading recommendations. For example, combining user ratings from platforms with different scoring systems can distort the overall ranking.

When 'data in the sources are mutually inconsistent...query answering becomes meaningless' (Lenzerini, 2002), making such data difficult to use for model training. In large-scale RS, this challenge is compounded by the reliance on vast datasets sourced from multiple repositories. Data federation allows queries across distributed sources without central consolidation, increasing coverage but introducing incompatibilities in structure, format, and meaning. Differences in schema design, controlled vocabularies, coding standards, and integrity constraints create both semantic and structural mismatches. The same attribute may be recorded in different formats, terms may have divergent scopes, and identifiers for the same entity may not align. Inconsistencies can occur at multiple levels: schema-level conflicts arise when field names match but refer to different concepts; value-level conflicts occur when records for the same entity differ in labelling or identifiers; and constraint-level conflicts result from uneven enforcement of validity rules, allowing data in one source that would be rejected in another. These mismatches can propagate through the RS pipeline, affecting training data consistency and influencing how outputs are prioritised within the

system. Although researchers have developed methods for ‘formally dealing with inconsistencies in databases, in particular for providing informative answers even in the case of a database that does not satisfy its integrity constraints’ (p. 240), such approaches are rarely applied at commercial scale. Large RS often address these issues through ad hoc transformations, local normalisation, or selective exclusion, without resolving the underlying misalignments. Without governance over vocabulary use, data representation, and integration rules, scaling through federation can increase volume while reducing coherence, producing recommendations shaped by integration artefacts rather than meaningful synthesis.

Data transformation modifies collected data to improve the accuracy of subsequent analysis, relying on ‘commonly-used tools that can serve many functions in quantitative analysis of data, including improving normality of a distribution and equalizing variance to meet assumptions and improve effect sizes, thus constituting important aspects of data cleaning and preparing for your statistical analyses’ (Osborne, 2010). Some data are discarded during this process through mathematical and statistical transformations, such as square root or log transformations used in RS evaluation metrics and loss functions. For instance, applying a log transformation to article view counts in a news recommendation dataset compresses differences between widely read and niche articles, potentially reducing the relative visibility of less frequent content within recommendation outputs. Data reduction, another pre-processing step, can significantly influence which content is considered relevant and therefore the quality of recommendations. While ‘eliminating data not relevant to the analysis

at hand—or extracting the data that are relevant—is usually the first, and arguably the simplest, form of data reduction’ (Namey et al., 2008), it can limit interpretability, conditioning how users engage with and make sense of recommendation outputs. Data discretisation, which converts continuous variables into categories, is particularly important for video-hosting platforms, as is seen in the next chapter. For instance, a RS dataset may require ratings such as ‘excellent,’ ‘good,’ and ‘bad’ rather than continuous numerical scores. Data discretisation is ‘one of the most effective data pre-processing techniques’ in data mining, yet ‘no distributed approach has been developed in the literature’ (Ramírez-Gallego et al., 2016). This is significant because large-scale RS datasets are often stored across multiple servers or data centres. Without a distributed method, transformation must occur on a single machine, creating bottlenecks and limiting scalability. Discretisation also imposes arbitrary boundaries that may not reflect underlying events in the dataset. To reduce these risks, it is important to approach discretisation with a clear understanding of the trade-offs it entails.

The pre-processing stages described above address missing data created by the methods used to clean, transform, or extract features from raw datasets, ensuring that such gaps are identified and managed before model training. Missing data in raw datasets is a common issue that can arise even before pre-processing begins. For instance, inconsistencies can occur when the same item is recorded in several different forms, such as US, USA, or United States. Techniques to address missing data include listwise deletion, where an entire row is removed if any values are incomplete. Imputation involves predicting

missing values based on other available data. Interpolation or extrapolation fills gaps using preceding or succeeding time points. These methods can introduce bias by adding unsupervised data that reflects statistical patterns rather than actual relationships between items. For this reason, missing data is ‘one of the most important statistical and design problems in research’ (Baraldi & Enders, 2010).

The discussion of data bias and provenance shows how these factors shape every stage of RS development, from data collection to model deployment. Provenance and lineage influence the extent to which systems are vulnerable to effects such as shrinkage, underfitting and overfitting, statistical discrimination and the trade-off between speed and thoroughness in decision-making. Bias or discrepancies in source data can propagate through these processes, altering what is recommended and to whom, and contributing to an epistemic agency gap. Building on this cultural framing, the next section examines how these factors interact with core quality dimensions of RS, including accuracy, diversity, serendipity and novelty.

Quality: Accuracy, reliability, and the limits of optimisation

While LF define how a RS is trained and what it optimises internally, quality measures address how the behaviour of a trained system is evaluated once it is deployed. The distinction is analytical rather than purely technical. LF operate within the system during training and are largely opaque to external observers, embedding optimisation priorities at the programming level. Quality

metrics, by contrast, assess externally verifiable aspects of system performance, such as accuracy, reliability, diversity, and stability, and are used to compare, audit, and regulate RS behaviour across contexts. The Quality section therefore examines how optimisation outcomes are interpreted, measured, and legitimised after training, rather than how they are produced within the model.

Quality in RS involves multiple interacting dimensions, commonly discussed as accuracy, diversity, serendipity, and novelty. These dimensions address both the technical alignment between predicted and observed user behaviour and the system's capacity to broaden the range of material a user encounters. Accuracy refers to the degree to which recommendations match user preferences as inferred from prior interactions. Diversity concerns the variety of recommended content types, reducing over-reliance on a narrow set of items. Serendipity introduces relevant but unexpected items that extend user exposure beyond predictable choices. Novelty prioritises unfamiliar content over repeated suggestions, although excessive novelty may reduce relevance. The relative weight of these elements varies across contexts, but together they shape user perceptions of recommendation value. Empirical work indicates that accuracy dominates RS quality research, while other measures receive less attention. Khusro, Ali and Ullah highlight the 'non-availability of benchmark dataset that can be used in evaluating a recommender system in a particular domain,' which limits cross-domain comparability (Khusro, Ali & Ullah, 2016). Bobadilla et al. note that 'research into recommender system quality measures has focused on accuracy,' defined as the correspondence between predicted and observed preferences, while 'an important lack of research into

reliability/confidence quality measures' remains (Bobadilla et al., 2018). These choices influence user trust and shape patterns such as homophily, where recommendations reinforce similarities in user networks. Homophily can increase satisfaction but may also limit exposure to diverse perspectives.

Quality is also shaped by how RS incorporate ethical considerations. Privacy, fairness, and transparency, especially in relation to legal compliance, can alter trust and engagement. A system may provide consistently high-quality recommendations to one segment of its user base while another segment receives lower-quality outputs (Ekstrand et al., 2018). Disparity in service levels can influence perceptions of fairness, shaping how users evaluate the system's credibility. Similarly, insufficient privacy safeguards such as extensive tracking without clear consent can lead users to reduce interaction or disable personalisation, indirectly lowering recommendation quality. RS quality interacts with broader platform attributes. In Netflix, which 'offer many more choices than humans can process, diminishing the quality of their decision-making abilities' (Ekstrand et al., 2018), the RS operates alongside interface design, catalogue breadth, and brand reputation to influence retention. On Spotify, discovery features coexist with personal playlists and curated content, meaning RS outputs are only one factor shaping engagement (Musosoup, 2025). The relationship between ethics and quality is therefore bidirectional. Ethical safeguards such as privacy and fairness influence trust, which affects engagement patterns and, in turn, the quality of recommendations. Statistical techniques used to optimise accuracy, diversity, serendipity, and novelty have implications not only for RS

performance but also for how ethical concerns such as bias, exclusion, or unequal service are addressed in practice.

The shrinkage effect is a statistical technique used in RS development that reduces variation in model predictions by drawing estimated values towards an overall average. This adjustment simplifies the model and reduces the risk of overfitting, a process discussed later in the chapter, in which the model becomes too closely matched to the training data. By constraining variability, shrinkage supports generalisation and allows the RS to perform more consistently across a wide range of inputs rather than being highly specific to the original dataset. In this sense, shrinkage can improve the stability and predictive reliability of recommendations, contributing to perceived quality. However, overuse of shrinkage can suppress distinctive patterns present in the data. This may reduce diversity or novelty in recommendations, which can diminish the system's capacity to expose users to less common or unexpected content. Techniques that push predictions towards central values prioritise dominant trends, potentially overlooking minority patterns or niche items. In some methods, 'some of the coefficients may be estimated to be exactly zero' (James et al., 2013), effectively removing certain variables from consideration. This can reduce exposure to items outside the mainstream and may limit opportunities for users to encounter material that would broaden their decision-making. Such effects can have both functional and ethical implications. Functionally, an RS may become overly concentrated on items that already receive substantial interaction, creating a self-reinforcing cycle in which popular items gain further visibility. From an ethical perspective, the narrowing of exposure can condition informational

diversity within recommendation outputs, prioritising well-represented content over less frequent items. The balance between the stabilising benefits of shrinkage and its potential to narrow recommendation scope is therefore central to how RS quality is defined and evaluated.

Underfitting and overfitting represent two contrasting points along the continuum of model complexity in RS. Both have a direct impact on performance and the overall value of recommendations. The distinction relates to the balance between simplicity and complexity in the model, with accuracy referring to how closely recommendations align with user needs, and quality encompassing broader factors such as relevance, consistency, and the ability to support informed decision-making. Underfitting can occur when ‘the discovered model allows for much more behavior than actually recorded in the log’ (Koehrsen, 2018), meaning the system predicts a wider set of actions than the data indicates, suggesting a poor fit to observed patterns. In such cases, the model overlooks important relationships in the data and produces imprecise or irrelevant recommendations. Overfitting, described as ‘a key problem in the supervised machine learning tasks’ and as ‘the phenomenon detected when a learning algorithm fits the training data set so well that noise and the peculiarities of the training data are memorized’ (Jabbar and Khan, 2015), occurs when the system is tailored too closely to the training dataset. While this can produce strong results on known data, it often limits the system’s ability to generalise to new inputs, leading to inconsistent or inaccurate outputs. Balancing the risks of underfitting and overfitting is central to RS design. A model that neither oversimplifies nor excessively complicates the data is more likely to sustain relevance and

accuracy (Van der Aalst et al., 2010). The process of achieving this balance involves trade-offs, where adjustments to model complexity can alter which recommendations are shown and how they are prioritised. These decisions reflect both technical judgement and contextual factors, producing outcomes that may differ from those generated in less constrained settings. The ethical implications of underfitting and overfitting extend beyond technical performance. An underfitted system, by offering inaccurate or overly generalised recommendations, can unintentionally steer user choices and limit agency. It may also emphasise existing biases in the data, reinforcing dominant narratives and reducing exposure to alternative perspectives. While overfitting can produce similar effects by memorising and amplifying dominant signals in the training data, underfitting does so through excessive generalisation that suppresses meaningful variation. This can limit informational diversity and contribute to uneven visibility across user groups. Overfitted systems can pose transparency challenges, as their reliance on memorised data patterns may produce recommendations that are difficult to interpret or justify, affecting user trust.

Statistical discrimination, in the context of RS, refers to the use of aggregated group characteristics to inform predictions about individual preferences. This can occur when a system draws on statistical patterns linked to demographics or other group-level features, rather than incorporating the unique attributes of each user. Such an approach may replicate existing patterns in the data, which can condition the range of content surfaced for different user groups. Research in medicine has shown that ‘statistical discrimination is a potential source of racial/ethnic disparities, and worthy of research’ (Balsa and

McGuire, 2001), indicating the relevance of this phenomenon beyond a single domain. While statistical discrimination can increase efficiency by simplifying prediction models, it may also compromise RS quality dimensions such as diversity and novelty. The use of group-level data raises questions about fairness and transparency, particularly in how recommendations are justified and interpreted by users. Duéñez-Guzmán et al. define it as ‘selecting social partners based not on their underlying attributes, but on readily perceptible characteristics that covary with their suitability for the task at hand’ (Duéñez-Guzmán et al., 2021). In RS contexts, this often means that models prioritise patterns that occur most frequently in the data, which can favour majority groups and replicate their preferences in recommendations. Models of statistical discrimination ‘do not attempt to capture the full complexity of racial, gender, and other types of discrimination in human societies, which likely involve a wide range of cognitive phenomena such as in-group favoritism, group entitativity, and more’ (p. 1). This simplification can produce an environment in which ‘limits to time, information processing ability, or the reliability of signals of partner quality...impact the accuracy of decision making, and produce higher bias’ (p. 2). As users have greater capacity to process information about others, ‘less discriminatory behavior results in higher payoff’ (p. 6), suggesting that more individualised and transparent recommendation processes could mitigate bias. Given its influence on both the quality and ethics of RS, statistical discrimination occupies a dual position in the field. It can support operational efficiency yet risk reducing informational diversity and reinforcing social inequalities. This relates closely to another important aspect of RS design: the dynamics of fast and long decision processes, which shape how recommendations are produced and evaluated.

Fast and long decisions in RS describe the trade-off between immediate user engagement and long-term satisfaction and retention. Fast decisions seek to capture interest quickly through recommendations that are immediately appealing, while long decisions focus on sustaining engagement over time by providing recommendations that are more diverse, accurate, and individually tailored. The balance between these approaches shapes RS functionality and influences overall quality by determining both the short-term and long-term effects of recommendations. Immediate relevance can increase early engagement and perceived usefulness, yet reliance on fast decisions alone can narrow the scope of content presented, reducing diversity and leading to disengagement over time. Long decision strategies, when well designed, can build sustained interaction by expanding exposure and refining accuracy, but they require greater investment in data analysis and content variety. This balance presents specific ethical challenges. A strong focus on fast decisions can limit informational diversity and reinforce existing preferences, potentially conditioning patterns of exposure by aligning recommendations with existing preferences, which may limit the range of perspectives encountered. Long decision strategies, while potentially enhancing diversity and depth, often involve more extensive data collection, raising privacy concerns. Decisions about where to position RS on this continuum therefore have implications for both user autonomy and the ethical standing of the system. Machine learning models used in decision-making can unintentionally reproduce patterns of discrimination, raising fairness concerns. Research into ‘discrimination-aware data mining’ and ‘fair machine learning’ aims to identify and mitigate such bias (Binns, 2018). Approaches

include comparing outcomes between protected and non-protected groups, assessing whether accuracy is consistent across groups, and exploring counterfactual scenarios to test the effects of removing or altering sensitive features. Achieving all fairness measures simultaneously is mathematically unfeasible in most contexts, which forces trade-offs between competing fairness definitions. Determining what constitutes legitimate grounds for different treatment is itself an ethical question that often depends on context. Barocas, Hardt, and Narayanan note that ‘there is no straightforward criterion that allows us to assess if a feature is morally valid ... this question must be debated on a case-by-case basis’ (Barocas, Hardt and Narayanan, 2023). For instance, in a hiring RS, gender may be excluded to prevent bias, yet removing it can reduce the model’s ability to account for legitimate job-relevant patterns if those patterns are statistically linked to gender. Such links might include career breaks associated with parental leave, participation in specific training programmes, or representation in industries with long-standing gender imbalances. Choosing between fairness criteria, such as demographic parity or equalised odds, depends on the outcome prioritised, and in practice no single approach can satisfy all fairness definitions at once.

The optimisation horizon refers to the time frame over which an RS is designed and evaluated, determining whether it is tuned for short-term or long-term objectives. Short-term optimisation relies on immediate engagement signals such as click-through rate or session duration, favouring already popular content and reinforcing narrow recommendation patterns. Long-term optimisation models delayed effects, such as sustained interaction with varied content or

gradual changes in user interests, but requires longitudinal data that often depends on persistent identifiers and probabilistic re-identification, raising privacy concerns. The temporal scope of optimisation is a design choice with direct implications for diversity, autonomy, and fairness. Short-term models can amplify popularity bias, while long-term models can increase surveillance risks. Hybrid approaches attempt to mitigate these effects by introducing diversity into fast-decision outputs or limiting the granularity of data used for long-term modelling. Such approaches require metrics beyond engagement, including diversity indexes and measures of epistemic range. Commercial incentives often draw systems toward short-term optimisation, as it is easier to measure, adjust, and monetise. Long-term optimisation can lock systems into expanded data collection practices, creating dependencies that are difficult to reverse. In both cases, the optimisation horizon functions as a design parameter that influences how content is prioritised within recommendation outputs.

Conclusion

This chapter has examined how the behaviour of a RS is shaped by the loss function that defines its optimisation target. The loss function establishes what the model treats as error and what it treats as signal, embedding priorities into the system before any recommendation is made. The selection of evaluation metrics then determines how performance is assessed and when training is considered complete. Metrics tied to clicks, watch time, or conversions align optimisation with those outcomes, influencing what the system explores and what it overlooks. The discussion then turned to data provenance, emphasising that the origins, labelling, and collection processes of the data constrain what the

model can represent. Historical biases, sampling gaps, labelling inconsistencies, proxy measures, imbalances, and missing contextual signals all shape the system's outputs unless they are explicitly addressed. Provenance information enables auditing and targeted intervention to mitigate these effects. The final part of the chapter analysed data processing within the RS pipeline, including extraction, cleaning, integration, transformation, dimensionality reduction, and handling of missing values. Each stage alters the statistical properties of the data and can remove, compress, or amplify particular patterns. Decisions such as discretisation, transformation functions, and shrinkage affect the system's ability to generalise, interact with risks of underfitting and overfitting, and influence the presence of statistical discrimination. These processes also interact with temporal optimisation strategies, including the balance between fast and long decisions, shaping how trade-offs related to diversity, privacy, and autonomy are configured within the system. Taken together, the definitions embedded in loss functions and evaluation metrics, the constraints set by data provenance, and the transformations applied during processing delineate key operational parameters of a RS. The next chapter applies this analytical lens to video-hosting platforms, examining how these design choices manifest in real-world recommendation flows.

Chapter 3. Recommender Systems: Case Study

The exploration of state-of-the-art recommender systems (RS) in the literature review chapter identified several contemporary techniques used by developers to train and evaluate algorithms. These methods may negatively affect citizens' ability to exercise their right to cultural participation and can also shape the quality of information access. In turn, this has implications for the quality of informed decision-making and for the conditions under which democratic participation takes place. While these risks are often described in abstract or technical terms, their impact becomes clearer when examined in specific social and political contexts. This chapter therefore investigates how RS influence linguistic diversity and political discourse through two case studies: YouTube in the United States and YouTube and Dzen in Russia. Analysing RS in both democratic and authoritarian environments makes it possible to observe not only how algorithmic practices constrain exposure to information, but also how political systems interact with these technologies to reinforce or challenge such constraints.

3.1 Methodology

This chapter applies the four elements of Technology, Economy, Culture, and Quality (Webster, 2014) to empirical case studies, shifting from theoretical framing to an analysis of how RS operate in practice. Results and interpretation are presented together and organised according to Webster's thematic dimensions to maintain conceptual continuity between empirical observations and their socio-technical implications. While this departs from a conventional

separation of results and discussion, the approach reflects the exploratory nature of the study and prioritises analytical coherence over procedural segmentation. Descriptive findings are therefore embedded within the analytical framework through which their relevance to democratic participation is assessed. To explore these four elements, the chapter draws on two case studies that examine patterns in RS outputs and consider how particular development strategies may contribute to them. The first case study focuses on YouTube in the USA and the content recommended to users during the four months preceding ‘Super Tuesday’ 2024, a key date in the presidential election primary cycle, when many states hold their balloting and frontrunners typically emerge (Guardian staff, 2024). Super Tuesday functions as a coordination point in the US electoral cycle, concentrating media attention, campaign activity, and public interest across multiple states within a short time frame. This convergence produces heightened volumes of politically relevant content and increased competition for visibility, creating conditions under which recommendation systems are likely to prioritise engagement-optimised outputs. Using this period as a benchmark therefore allows the analysis to observe recommendation behaviour under intensified political salience rather than during routine news cycles. The second case study analyses discourse around the war in Ukraine on YouTube and the Russian video-hosting platform Dzen during the four months preceding the 2024 Russian presidential elections. In contrast to the US case, where election-related content dominated recommendations, Russian political discourse across both state-aligned and oppositional sources framed societal issues through the lens of the war in Ukraine. Pro-government narratives present the war as justification for domestic challenges or as the reason for national unity, while oppositional voices

use the war to critique the regime and question its legitimacy (Alieva, Kloo & Carley, 2024). As there is no viable electoral competition, the war served as the dominant reference point through which presidential discourse is constructed and contested in Russia (Oleinik, 2024). Placing this case alongside the US election context highlights how RS operate under both competitive and non-competitive political conditions, allowing a comparative view of how different environments shape the circulation of discourse. Both case studies analyse the linguistic content of recommendations to show how RS shape not only which items are promoted but also the lexical characteristics of the discourse they contain, creating gaps in linguistic diversity. Each case links these patterns to the design of RS, where optimisation techniques such as loss function constrain item variety.

This chapter uses linguistic analysis because the effects of RS on democracy can be observed most clearly in the language that users encounter through recommended content. The method is applied to the case studies to evaluate how RS outputs structure political discourse. Measures such as lexical diversity, sentiment, and density are treated as indicators that suggest the breadth of perspectives available and the conditions under which citizens can participate in democratic debate. Political discourse is mediated through words, and shifts in vocabulary, framing, and tone directly influence how citizens interpret events, positions, and actors. By examining lexical diversity, sentiment, and density, it is possible to measure the range and character of political expression made available by recommendation outputs. These linguistic signals do not evaluate the truth of statements, nor do they assess factual accuracy.

Instead, they reveal structural properties of discourse that condition how individuals can participate in civic life. Lexical diversity can be understood as a proxy for the breadth of perspectives in circulation. A system that repeatedly presents content with low vocabulary variation constrains interpretive flexibility, making it harder for users to engage with alternative frames of reference. Vocabulary size captures the scale of linguistic resources required for comprehension, which can either empower or restrict participation depending on users' capacity to process complex texts. Lexical density measures the concentration of content words relative to function words, indicating whether messages are conveyed in a repetitive and simplified form or through more information-rich expressions. Together, these metrics provide a systematic way to track how RS configure the linguistic environment of political debate. Sentiment analysis complements these measures by showing how affective framing shapes discourse exposure. RS frequently optimise for engagement by amplifying emotionally charged material. This introduces predictable biases in the distribution of positive or negative tones in discourse across different groups. By capturing shifts in sentiment balance, such analysis identifies how emotional reinforcement may lock users into cycles of anger, fear, or approval, limiting opportunities for balanced evaluation. Previous research across linguistics, political communication, and information systems has used these metrics to document how online platforms influence language use and participation (Chilton, 2004; Nation, 2006; Ribeiro et al., 2020). Applying them here creates a bridge between technical studies of RS and theories of democratic discourse. These indicators provide accessible entry points: they reduce vast amounts of textual data into interpretable signals about inclusivity, variety, and emotional

framing. The case studies, therefore, use linguistic analysis not as an end in itself, but as a method to evaluate whether RS outputs widen or narrow the conditions for civic participation. In addition to aggregate lexical metrics, the analysis combines distant reading with targeted close-reading diagnostics to interpret observed patterns in the data. Frequency distributions, recurrent lexical bundles, and coarse topic signals were used during analysis to identify which words and phrases contributed to repetition and divergence between vocabulary size and lexical diversity. These diagnostic procedures were applied to contextualise metric outcomes, for example where low Type-Token Ratios coincided with large vocabularies, by examining whether repetition was driven by a narrow set of high-frequency terms, proper names, or formulaic framings. This approach was not intended as a full topic modelling or discourse-analytic study. Rather, it served as an interpretive aid to confirm whether observed metric patterns reflected structural properties of recommendation outputs rather than artefacts of preprocessing or aggregation. The emphasis therefore remains on lexical indicators of exposure structure, supported by selective textual inspection, rather than on exhaustive thematic classification.

The structure of each case study provides a framework for analysing the impact of RS on the linguistic characteristics of the online content that citizens are exposed to. The *Technology* section explores how algorithms shape lexical diversity, temporal patterns, and personalisation, examining how algorithmic design relates to patterns of linguistic exposure that are analysed through the concept of epistemic agency. These metrics reflect how RS modulate user exposure to language variation, time-sensitive content, and individual tailoring,

all of which shape access to political discourse. The *Economy* section emphasises commercial drivers such as demographic targeting and platform monetisation strategies, examining how linguistic features such as lexical density and the formation of echo chambers support advertising goals and influence epistemic agency. These elements reveal how economic optimisation strategies indirectly shape the informational environment by privileging content that maximises engagement. The *Culture* section examines vocabulary size, cultural sentiment, and narrative evolution, highlighting how generational variations correlate with democratic discourse. These indicators are used to examine how cultural narratives vary across age groups within the analysed datasets, highlighting differential patterns of visibility in political communication. The *Quality* section analyses sentiment balance, consistency, stability, and text breadth, addressing ethical considerations around demographic-based curation and its effect on user's agency. These signals are used as proxies for assessing whether recommendations promote emotionally coherent, thematically diverse, and demographically equitable content exposure. This analysis covers how RS influence online discourse across various platforms and user groups.

The first case study collects content recommendations from YouTube in the USA between December 2023 and March 2024. This period is selected to capture the period leading up to 'Super Tuesday,' which was held on 5 March 2024 (Guardian staff, 2024). Data gathered from these recommendations was analysed using methodologies from previous research on the partisan and radicalising effects of YouTube in USA (Alfano et al., 2021; Covington, Adams & Sargin, 2016; Davidson et al., 2010; Kaiser & Rauchfleisch, 2020; Ledwich &

Zaitsev, 2019; Munger & Phillips, 2022; Ribeiro et al., 2020; Schmitt et al., 2018; Toderici et al., 2010; Tufekci, 2018; Zhou, Khemmarat & Gao, 2010). Building on these works, this case study analyses the top recommendations using content analysis techniques similar to those of Zhou et al., Ribeiro et al., and Munger and Phillips (Zhou, Khemmarat & Gao, 2010; Ribeiro et al., 2020; Munger & Phillips, 2022). The research utilises a manual large data collection method similar to that of Schmitt et al., and incorporates demographic factors as in Davidson's work (Schmitt et al., 2018; Davidson et al., 2010). By applying large-scale data analysis techniques similar to Covington and Toderici, the study offers descriptive insights into patterns of linguistic diversity observed in recommended content on video platforms (Covington, Adams & Sargin, 2016; Toderici et al., 2010).

The second case study focuses on both YouTube and Dzen in Russia during the same period, from December 2023 to March 2024. This is the time leading up to the presidential elections in Russia held from 15 to 17 March 2024. Due to the war in Ukraine and its impact on Russian society, discourse about its implications became a primary topic before the election. The case study is divided between the two platforms because the Russian government bans and restricts any anti-war content on the Russian video-hosting platform Dzen, to the extent that posting anything deemed anti-war is a prosecutable felony in Russia. YouTube, on the other hand, bans all pro-war content in the Russian language for the Russian region. Although this study combines results from two different platforms, the distinction reflects how each platform enforces content rules: YouTube's restrictions result in the dataset containing primarily anti-war material,

while Dzen's restrictions result in the dataset containing primarily pro-war material. The chosen time frame may not represent long-term trends, but it captures a concentrated period of heightened political activity in the run-up to the election, when RS outputs are likely to be more visible within public communication environments. However, the case study focuses primarily on the state of the political environment and the quality of cultural participation driven by the recommendations. These elements become particularly vulnerable during election campaigns and periods of agitation in the news. These issues are also of interest to advertisers, who are key users of the video-hosting platforms and whose goals serve as primary economic drivers for the developers. The analysis of the case studies focuses on linguistic variation between political discourses, using measures of vocabulary diversity, language complexity, sentiment patterns, and demographic-specific content trends. The computational linguistics metrics are used to reflect on the emerging gap in users' access to discourse, which manifests through measurable linguistic differences shaped by the RS development methods described in Chapter 2. To examine how recommendations are generated for new users, the analysis makes use of the cold start problem, previously identified as a primary challenge shaping developers' approaches to RS architecture. This is addressed through sock puppets (SP), which are new accounts with no history (Liu, Wu & Resnick, 2024). RS struggle to recommend content to such blank accounts, so platforms employ methods discussed in Chapter 2, such as using external information about the accounts or using hybrid systems to create recommendations. The use of SP limits the influence of prior user preferences or interaction history, allowing observation of recommendation patterns under cold-start conditions. In total,

sixty SP were used in this research, divided into three gender categories (male (M), female (F), rather not say (N)) and five age groups (18-24, 25-34, 35-44, 45-54, 55-64). Fifteen SP were created for each scenario, marked with identifiers designed for clarity of reference: USB (US Blue) for YouTube — Democratic Party, USR (US Red) for YouTube — Republican Party, RUB (RU Blue) for YouTube — anti-war in Ukraine, and RUR (RU Red) for Dzen — pro-war in Ukraine. This structure is driven by the internal categorisations of the users for the advertisers on Google platforms using age targeting, where ‘valid age ranges are 18-24, 25-34, 35-44, 45-54, 55-64, 65 or more, and Unknown’ (Google Ads, 2025). This chapter does not include the group ‘65 or more’ due to the broadness of their data, which could have introduced significant noise to the dataset. It also excludes ‘Unknown’ as creating a new Google or Dzen account requires age identification.

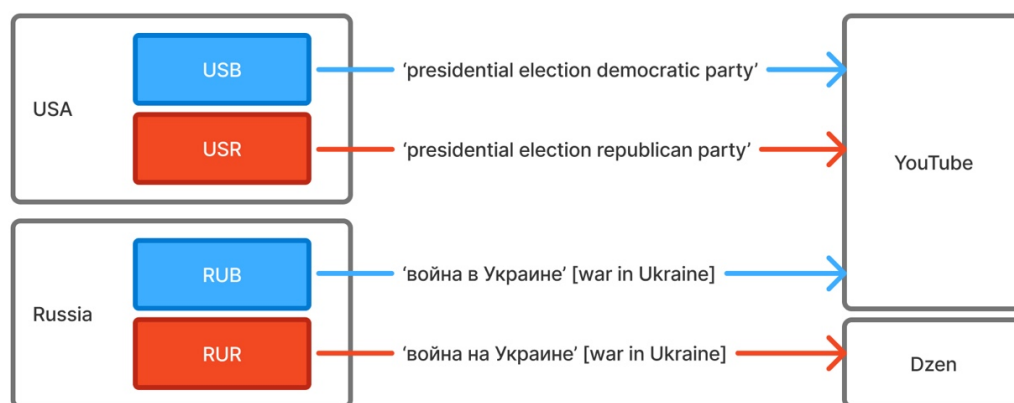


Figure 1 Sock puppet categories, associated queries, and platforms.

Each SP browses both YouTube and Dzen through a chain of recommended watch-next videos by selecting the top recommendations, similar to method used in Faddoul, Chaslot & Farid’s research and based on the fact that ‘70% of watched content on YouTube is recommended content’ (Faddoul, Chaslot & Farid, 2020; Solsman, 2018). The first video in the chain is chosen

based on a specific phrase typed in the search: USB (YouTube, US) – ‘presidential election democratic party,’ USR (YouTube, US) – ‘presidential election republican party,’ RUB (YouTube, Russia) – ‘война в Украине’ [war in Ukraine], RUR (Dzen) – ‘война на Украине’ [war in Ukraine]. The distinction in spelling between RUB and RUR comes from the cultural legacy of the USSR. The grammatically correct usage in Russian uses ‘в’ [in] with country names, such as ‘в Ирландии’ [in Ireland]. However, the use of ‘на’ [on] reflects a USSR-era perception of Ukraine as a region of Russia, similar to how regions within Russia are referred to: ‘на Урале’ [in the Ural] or ‘на Алтае’ [in the Altai]. Since Russia’s invasion of Ukraine in 2014, the change in this linguistic practice has become highly politicised and part of the new social-cultural trend, representing a ‘linguistic confirmation of its [Ukraine’s] status as a sovereign state’ (Graudina, Itskovich & Katlinskaia, 2001; Devlin, 2016), while in Russia the older form continues to be used to reinforce a perception of Ukraine as a subordinate region. The selection of these short phrases is used to seed the recommendation process within a specific topical context associated with each country and political position. This allows the platform’s algorithm to drive subsequent recommendations without interference, showcasing the gathered data and underlying algorithmic influence on citizens’ information journeys. These algorithmic patterns shape the informational context within which users encounter politically relevant content.

Using this approach, the links to the first 10 recommended videos were collected every three days for each SP. Over four months of data collection, USB, USR, and RUB collectively watched 17,964 videos on YouTube, while RUR

watched 4,883 videos on Dzen. The sampling interval reflected three considerations. The first was Sampling Theory (Lohr, 2021), where a three-day interval provided a balance between variance and bias in time-series data. Daily collection risked capturing short-lived anomalies such as sudden spikes in recommendations triggered by algorithmic testing or ephemeral trending content. Weekly sampling risked missing the influence of time-sensitive events, such as political speeches or breaking news, which can alter RS outputs for several days before returning to prior patterns. The second consideration was data overload, where reducing the total number of collected recommendations helped to limit overfitting during analysis. High-frequency collection would have increased duplication from the same content creators, inflating similarity scores and masking less frequent but significant recommendation patterns. The third was quality of analysis (Raimondi, Skene & Emmerich, 2008; Ghezzi & Guinea, 2007; Murthy, 1967). A three-day interval created clearer separation between shifts in topics, allowing patterns to be traced without excessive interference from targeted repetition or rapid content turnover. This spacing supported both manual review and automated analysis by ensuring that each snapshot reflected a stable stage of the recommendation environment. This interval was, therefore, sufficient to capture changes in recommendation patterns without introducing unnecessary noise, while also enabling efficient use of computational resources over the four-month study period. In the results, measurements are aggregated at the monthly level to foreground sustained patterns in recommendation outputs and to reduce short-term volatility arising from transient events, algorithmic testing, or episodic content bursts. 3-day variation is present in the underlying dataset but treated as

background fluctuation relative to the longer-term exposure patterns examined in the case studies.

Each account was run in a separate VMware virtual machine to create a 'blank' working environment for each SP. The virtual machines were operated on Ubuntu 22, a free Linux-based OS with minimal computational requirements. Each VM was connected to the internet through the Tor network. This approach ensured that recommendations were not influenced by IP addresses, reducing potential bias. For each day when data was collected, a new Tor circuit was generated for each SP. Every three days, data was manually gathered using each SP and stored in lists containing links to the video content. These lists were analysed for duplicates to optimise the data that required downloading and processing. The updated lists were then used to download videos from YouTube and Dzen in audio format using the open-source tool yt-dlp, which is commonly employed by researchers to download audio and video files (Coats & Laippala, 2024). The audio files were transcribed using the open-source tool Whisper, developed by OpenAI, using medium-sized models for both English and Russian transcription. Transcriptions were generated from the downloaded audio tracks of videos as retrieved via yt-dlp. Platform-inserted advertisements that are dynamically served by the hosting platforms are generally not included in these audio streams and therefore do not enter the transcripts. Creator-embedded advertising segments, such as sponsorship messages or promotional reads integrated into the video content, may be present in some transcripts. These segments were not explicitly removed during preprocessing. However, such content typically follows repetitive and formulaic language patterns and

constitutes a small proportion of total tokens, and is therefore unlikely to materially affect aggregate lexical metrics at the corpus level. Any residual influence is treated as part of the broader communicative environment encountered by users rather than as a separate analytical category. The resulting text files were analysed for language to ensure that the corpora consisted only of English for USB and USR, and Russian for RUB and RUR. For instance, the proportion of English-language videos for USB and USR was approximately 4692 (93%) and 4743 (94%), respectively. During data collection on YouTube, the cold start problem sometimes caused the algorithm to suggest videos in the language of the country where the IP connection appeared to come from, especially after the fifth or sixth recommendation. For example, when the IP connection appeared to come from Thailand, the platform recommended Thai-language videos, which were not relevant to the research. To exclude unrelated content, a language filter was applied. The language distribution of videos for RUB was: 2281 (42%) in Russian, 2169 (39%) in Ukrainian, and 1088 (19%) in other languages. The significant amount of Ukrainian-language content is due to Google's algorithm not distinguishing between Russian and Ukrainian content. Many Ukrainian creators produced content in Russian before the escalation of the war in 2022. After this, many shifted to creating content exclusively in Ukrainian, reflecting broader linguistic changes in the region mentioned previously (Центр контент-аналізу, 2023; VisitUkraine, 2023; Orobchuk, 2024).

Using the transcribed videos, several corpora were created for analysis. For each USB, USR, RUB, and RUR, the following were developed: overall corpora for each month (December 2023, January, February, and March 2024),

combined months (December 2023 + January 2024 and February + March 2024), corpora for all five age groups, and corpora for each gender. These corpora were analysed using Python, where spaCy (en_core_web_lg) was used for English and PyMorphy2 with Razdel for Russian language processing. The preprocessing phase included removing non-alphabetic characters and applying Gensim's standard stop word list. This was supplemented with a custom list, based on frequently occurring words that could affect the analysis, such as 'is,' 'they,' 'has,' 'were,' 'there,' 'which,' 'would,' 'could,' etc. The same approach was applied to the Russian corpora, which required a separate file containing stop words and another one with war-related terms to support sentiment analysis. This was necessary due to the specific terminology used by Russia to describe the conflict and the large volume of data. The war-related terms included: 'война' [war], 'конфликт' [conflict], 'вторжение' [invasion], 'оккупация' [occupation], 'спецоперация' [special operation], 'сво' [SMO (Special Military Operation)], 'боевые_действия' [combat operations], 'военные_маневры' [military manoeuvres], etc. The analysis focused on lexical richness metrics, including Type-Token Ratio (TTR), Vocabulary Size, and Lexical Density, to measure vocabulary diversity and content word distribution. Vocabulary counts are based on tokenised word forms after lowercasing, removal of non-alphabetic characters, and stop-word filtering using a combination of standard and custom lists. Named entities were not explicitly excluded and therefore contribute to vocabulary size where they survive filtering. Multi-word expressions were treated as separate tokens rather than as compound units. For Russian-language corpora, morphological normalisation was applied through PyMorphy2 lemmatisation, with acronyms and selected domain-specific terms preserved.

For English-language corpora, tokenisation relied on standard preprocessing without additional lemmatisation. TTR calculates the ratio between the total number of unique words and the total number of words in a text. It is a 'useful index of lexical diversity independent of sample size, and it has been used in this function in a wide range of studies' to indicate vocabulary richness where a higher ratio suggests more varied word usage while a lower ratio indicates more word repetition (Richards, 1987). For sentiment analysis, BERT models tailored to each language were used: a multilingual BERT model for English and RuBERT for Russian texts. These provided detailed measurements of positive, negative, neutral, and compound sentiment. The results of the analyses were compiled into CSV files, enabling convenient comparison. The quantitative analysis in this chapter is descriptive and exploratory. It reports observed differences and temporal patterns in linguistic metrics across recommendation outputs, rather than relying on inferential tests of statistical significance between groups or making generalisable causal claims. For this reason, the chapter does not apply inferential statistics or significance testing, and reported averages should be read as indicators of patterns in this dataset rather than as validated effect estimates. Transcript length varies across corpora due to differences in video duration, recommendation volume, and platform-specific content formats. To reduce instability arising from very short texts, analyses are conducted on aggregated corpora (monthly, demographic, and platform-level) rather than on individual transcripts. Token counts are retained for each corpus and used to contextualise lexical metrics, particularly when interpreting divergence between Type-Token Ratio and vocabulary size. As a result, observed differences are read as properties of aggregated exposure environments rather than of isolated texts.

While transcript length is recorded and considered in interpretation, the analysis does not normalise metrics by fixed token windows, which may introduce residual length effects in comparisons across platforms and time periods. Claims about recommender system behaviour in this chapter refer to observable variation in recommendation outputs under controlled exposure conditions rather than to internal system adaptation or learning at the level of individual user profiles. Given the cold-start design of the sock puppets, minimal interaction signals, and the use of rotating network locations, the analysis does not attribute changes in recommendations to personalised optimisation or dynamic user modelling. Instead, observed differences are interpreted as properties of exposure environments shaped by platform-level ranking logics, content availability, and contextual signals, without inferring adaptive intent or system responsiveness beyond what can be directly observed.

There are several limitations in the applied methodology that require attention. First, the choice of only two video platforms, YouTube and Dzen, may not fully represent the broader landscape of RS dealing with politically influential content. This chapter uses these platforms as examples to identify common trends and patterns that emerge when analysing how algorithms operate with data while forming epistemic gaps. Second, the manual selection of Russian war-related terms could have introduced researcher's bias. However, this should be a minor issue, as the majority of the content gathered is explicitly related to the war. Third, the quality of the collected data could vary due to the constant updates and fine-tuning of algorithms on these platforms. This was mitigated by restricting data collection to a four-month period. Nonetheless, technological

updates during this time may have introduced transient biases that could affect word frequencies and user recommendations, particularly during the Christmas and New Year holidays. This is especially relevant to SP due to the cold start problem as previously noted. To address this, the case study design allowed RS to tune themselves during December to adapt to each SP, generating more targeted recommendations over the following three months. Fourth, the Python tools selected for corpus analysis may not be optimal for capturing idiomatic or cultural expressions specific to political discourse. This limitation places greater responsibility on the interpretation of data, potentially introducing researcher bias. Therefore, the analysis primarily adopts a statistical perspective, enabling clear interpretation due to the specific choice of topics. The analysis does not formally distinguish between vocabulary classes such as named entities, technical terminology, or rhetorical repetition, and therefore treats vocabulary size as a coarse indicator of linguistic demand and repetition rather than as a direct measure of knowledge exchange or informational quality. The analysis also does not explicitly track the replication of identical recommended items across SP. While duplicate recommendations were identified during data collection to optimise downloading and processing, cross-account overlap was not treated as an analytical variable. This design choice reflects the focus of the study on the linguistic characteristics of exposure as encountered by individual user profiles, rather than on platform-wide convergence or content circulation patterns. Tracking replication across accounts would have enabled additional insights into the degree of homogenisation or synchronisation in recommendation outputs, but would have required a different analytical framing oriented towards network-level or distributional analysis rather than exposure-

level linguistic conditions. Finally, ethical considerations related to the use of SPs, including anonymity, non-interference, and data privacy, are discussed in relation to existing literature (Sullivan, 2014; Gera & Singh, 2015; Bu, Xia & Wang, 2013; Paterson, 2024). As these case studies involved only passive data collection, did not generate new interactions, and had no impact on platform users, ethical approval was not sought for this chapter.

The analysis does not attempt to separate the influence of recommender ranking mechanisms from temporal changes in the underlying pool of available content. As a result, shifts in linguistic patterns may reflect changes in content production, topical salience, or platform moderation practices alongside recommendation effects. The findings therefore describe how users are exposed to language through recommendation outputs at specific moments in time, rather than isolating algorithmic selection effects from content supply dynamics. The analysis adopts an external and observational position, examining recommendation outputs rather than the intentions of platform designers or content producers. The researcher does not participate in content creation or interaction and approaches the material as a linguistic artefact shaped by system-level selection mechanisms. Linguistic competence in both English and Russian enables direct analysis of discourse without reliance on translation, reducing interpretive distortion while increasing responsibility in qualitative interpretation. Close reading is used selectively to contextualise quantitative patterns rather than to infer authorial intent or platform objectives. This positional stance reflects a focus on exposure environments and structural tendencies rather than normative judgement of political positions.

3.2 US Case Study

During the period of this case study the sitting president in the United States was Joseph Biden, who was the Democratic Party's nominee and planned to run for a second term. The Republican Party had several nominees, including Nikki Haley, Ron DeSantis, and former President Donald Trump (BBC, 2024). The chosen period and search phrases for the SP were designed to capture discourse around the presidential elections from the perspectives of both the Democratic (USB) and Republican (USR) parties, following the algorithm's recommendations. The majority of the recommended discourse during this time centred on the legal proceedings involving former President Trump. By contrast, President Biden's position within the Democratic Party remained stable, as he faced no significant internal challenges and retained full support from its leadership (Pathe, Krieg & Bradner, 2024). This overlap, where Democrats largely discussed Trump's legal cases while Republicans promoted his candidacy, created a rare alignment that makes the timeline well suited for analysing how algorithmically recommended content was associated with differing vocabularies, sentiment patterns, and narrative focus across political discourses. The analysis provided offers insights into political discourse by highlighting patterns of algorithm-mediated cultural participation that are associated with linguistic differences between political party-aligned discourses.

Technology

From December to March, USB's average TTR was 0.058 and USR's is 0.052, with USB higher in December, January, and February, and lower in March.

Vocabulary size also varies by month, with USR exceeding USB in December, January, and February. In March, they are almost identical: USB 35,470 and USR 35,435. This similarity in size coincides with a difference in TTR: USB 0.0419 and USR 0.0480.

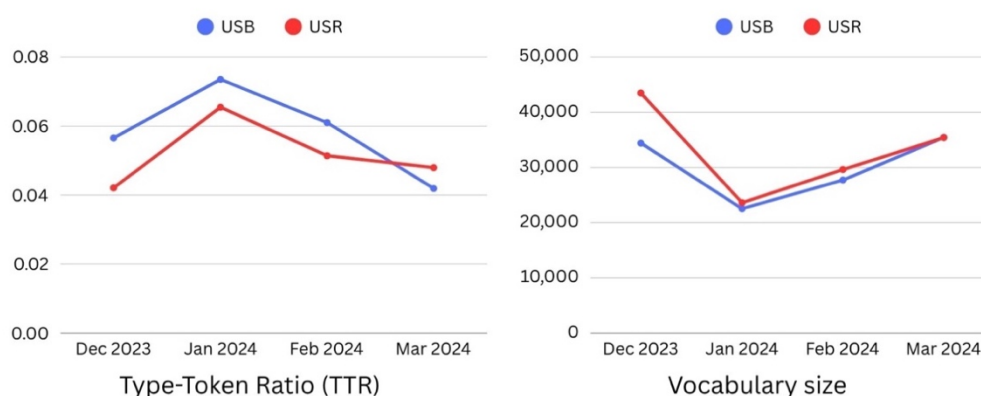


Figure 2 TTR and vocabulary size for USB and USR.

The temporal aggregation therefore suggests that USB content was lexically more varied than USR's for much of this period. However, when TTR is recalculated for demographic subgroups, the pattern reverses, with USB consistently showing lower TTR values than USR. This is not a contradiction but a result of how aggregation functions. A small number of high-variation contexts within USB disproportionately raise the monthly averages, even though all subgroup scores remain lower than USR's. For example, when a major news story breaks, USB recommendations may cover it from many angles, producing a rich overall vocabulary. Yet within each demographic subgroup, the coverage tends to recycle limited and targeted language, keeping the subgroup vocabulary much less diverse. A clear illustration comes from 2025, when twenty-two Democratic politicians shared an identical script on social media and recommendation systems redistributed it across all age groups, embedding the

same message in each subgroup’s targeted stream (Sky News Australia, 2025). This uniformity reduced lexical diversity inside every subgroup. However, the overall USB aggregate still appeared higher because occasional high-diversity events inflated the monthly average, masking the low-variation patterns revealed at subgroup level. This highlights a methodological limitation of relying on aggregated measures alone, since they can obscure the distributional patterns that appear across groups.

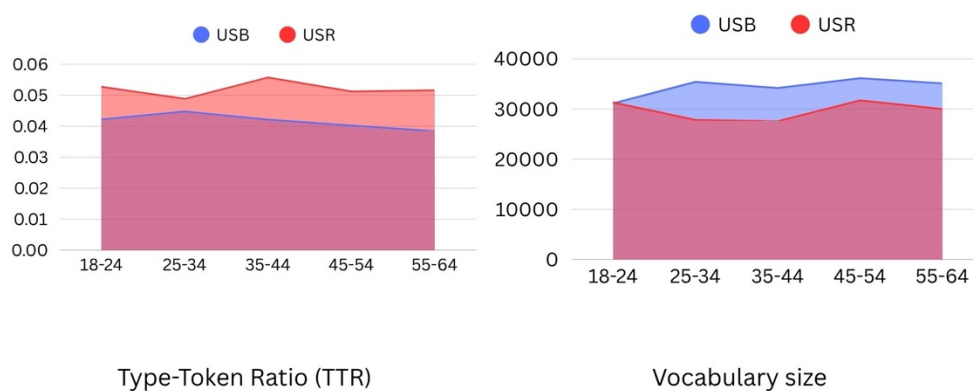


Figure 3 TTR and vocabulary size by age groups in US case study.

As discussed in Chapter 2, this is an example of the evaluation metric fallacy, where a global metric obscures how individual segments are actually targeted. LF optimised for overall engagement can converge on patterns that perform well on aggregate while exhibiting uneven distributional effects across groups. In this case, monthly TTR trends suggest greater lexical diversity for USB, yet the subgroup analysis reveals that the majority of its audience experiences less lexical variation than USR’s audience. This discrepancy does not come from a single subgroup being exceptionally diverse but from occasional high-variation events that raise the overall USB average while leaving most subgroup profiles low in diversity. The aggregation hides these disparities because the LF’s optimisation is blind to distributional equity, where it seeks to

maximise the chosen performance metric across the whole set, not to ensure balanced outcomes for each subgroup. This distinction between temporal and structural perspectives reflects a key methodological risk. Aggregated temporal measures capture short-term variation, often driven by discrete events, trending topics, or sudden influxes of novel vocabulary. Subgroup measures reveal persistent lexical tendencies that remain stable across audiences and sentiment contexts. The fact that USB's subgroup-level TTR is uniformly lower suggests that its lexical variety is episodic rather than systemic. In RS terms, this is consistent with an optimisation strategy where the model achieves its LF target by concentrating variation in high-impact episodes, while maintaining a more uniform and repetitive discourse for most of the audience. For studies of RS and political discourse, the implication is that momentary spikes in diversity lift the aggregate temporarily but are not assimilated into the wider discourse, so they mask the longer-term constraints that dominate the linguistic environment. A system may appear balanced in aggregate evaluation but still create asymmetries in exposure at the group level. In this sense, the monthly-versus-subgroup discrepancy illustrates how optimisation objectives interact with audience segmentation, highlighting differences between aggregate and subgroup-level patterns of linguistic exposure.

Two complementary constraints emerge in the lexical environment shaped by RS: oversimplification through low lexical diversity, which funnels users into filter bubbles, and overload through excessively large vocabularies, which limit comprehension. Both restrict democratic participation, one by narrowing interpretive flexibility and the other by raising barriers to entry. This

divergence between temporal and subgroup patterns becomes even more significant when absolute vocabulary sizes are considered. Nation estimates that 98% coverage in such texts can be achieved with about 9,000-word families, whereas, for instance, the USB corpus required 48,218 families, more than five times larger (Nation, 2006). While the primary analyses report vocabulary size in terms of token- and lemma-based counts, word-family estimates were additionally computed using established BNC/COCA family mappings to enable a more meaningful comparison with Nation. The comparison is therefore intended to reflect order-of-magnitude differences in lexical demand rather than strict equivalence between measurement units. The recommended content in this period therefore imposed substantial lexical demands on the user, far exceeding those needed to follow a novel or newspaper. These vocabulary sizes are typically associated with upper secondary or early tertiary-level texts, placing political discourse in a range that may be inaccessible to many users. In this way, RS generate discourse that simultaneously narrows variety while demanding advanced comprehension, conditions that may push significant parts of political communication beyond the accessible range for some citizens. At the same time, society produces a constant stream of complex and information-rich discourse, and RS, in seeking to meet their success metrics, respond by narrowing what is recommended.

Given the earlier finding that USB's higher monthly TTR is not reflected across subgroups, the reduced TTR in its overall profile suggests less linguistic variety in USB content, which may reflect a narrower range of expressions used across recommendations. Lower TTR values often correlate with texts written for

lower educational levels, where language tends to be more repetitive and restricted (Lu et al., 2017; Richards, 1987). This pattern is consistent with the reinforcement of echo chambers described in prior studies of online discourse, yet reduced lexical diversity can also contribute to comprehensibility across educational divides by lowering the linguistic threshold required for engagement (Baharmand, 2023). On the other hand, USR has higher TTR and broader linguistic variation, which suggests exposure to a wider range of content, minimising algorithmic pressure on users' perspectives. RS exposure patterns observed for USB align with findings that reduced lexical diversity is associated with constrained epistemic agency, although this analysis cannot isolate algorithmic effects from content availability or production practices (Brugnoli et al., 2019; Borah, Mihalcea & Pérez-Rosas, 2025). Such repetition reflects what Chilton describes as lexical framing, where consistent use of particular phrases activates fixed interpretive schemas and reduces openness to alternative viewpoints (Chilton, 2004). This narrows interpretive flexibility and limits opportunities to engage with competing frames of reference. While such cohesion could be consistent with a unified community narrative, it also constrains discursive openness and shapes political understanding within a fixed lexical boundary. This linguistic gap reflects patterns consistent with those documented in the literature on the influence of RS in shaping political discourse on YouTube.

The cold start problem amplifies these challenges, as new accounts often receive generalised recommendations until sufficient interaction data is collected. RS often default to generic or mainstream content, and persistent

negative sentiment may illustrate the system's limited adaptability to nuanced content needs. Other reasons for this behaviour are discussed later in Economy section. The system's architecture limits the data it can operate with by relying primarily on surface-level behavioural proxies such as clicks or watch time. It excludes richer contextual or historical user signals. This restricts the algorithm's ability to differentiate between nuanced content types and narrows the scope of recommendation. Sparse datasets prioritise dominant trends (Thompson et al., 2022). They sideline minority or less engaging perspectives, measured by clicks, watch time, or shares, and reinforce homogenised linguistic patterns. To address these issues, researchers focusing on fairness, diversity, and user autonomy in RS design prioritise content diversity (Varshney, 2020; Bonicalzi, De Caro & Giovanola, 2023; Lathia et al., 2010; Kunaver & Požrl, 2017; Castells, Hurley & Vargas, 2022). Hybrid systems, combining collaborative filtering and content-based approaches, can utilise external data to overcome cold start limitations, which improves the recommendations (Felfernig et al., 2013; Adomavicius & Tuzhilin, 2005). In theory such interventions can alter exposure patterns by increasing lexical diversity and reducing homogeneity in recommendations, although their effects on political discourse remain outside the scope of this analysis (Bakshy, Messing & Adamic, 2015). However, this approach is not a solution when applied to political discourse. Unlike recommending something on a marketplace, where introducing external biases might improve product recommendations, applying similar methods in RS which operate with political content risks compromising the neutrality of the platform. Introducing external bias to force diversity, while it may achieve symmetry in appearance, disrupts the system's inherent balance, masking rather than solving the underlying

issues. Political discourse requires solutions that empower diverse participation without computational intervention, ensuring platforms support genuine, unmanipulated representation of different opinions. USB displays significant sentiment instability, with negativity peaking in January (0.505) before bouncing back in February (0.361) and March (0.441). These drastic fluctuations suggest that recommendations related to Democratic Party discourse respond more strongly to external events, possibly reflecting the recollection of the events of 6th January 2021 during the Trump's trial (Brooke Singman, 2024). In contrast, USR demonstrates more stable sentiment patterns with negativity in January (0.467) and March (0.466). The stability in USR's sentiment indicates a more consistent curation strategy, reducing algorithmic fluctuations in content tone.

Temporal metrics also show notable changes in vocabulary size and TTR. USB's vocabulary size increased significantly in March (35,470), after dropping down in January (22,520). This could be consistent with an algorithmic attempt to diversify content gradually or a response to external events that prompted a temporary shift in recommendations. From January to March, USR vocabulary size rises from 23,610 to 35,435. Across December to March, it falls then rises. USB's TTR is lower than USR's only in March. In the other months USB's TTR is higher. These findings are consistent with the view that RS are associated with temporal variation in linguistic diversity, shaping patterns of exposure to evolving or repetitive discourse, although this analysis cannot establish effects on epistemic agency. Processing large-scale temporal data creates significant scalability challenges for RS due to the need of significant computational recourses. Platforms such as YouTube require continuous processing of

extensive datasets to account for shifts in user behaviour and external events which might affect the recommendation, such as advertisement campaigns. For instance, the content volatility observed in USB during January may indicate algorithmic limitations in responding to dynamic political discourse in real-time. Computational constraints can result in delayed or homogenised recommendations, limiting the algorithm's ability to adapt dynamically. This scalability issue, as discussed in Chapter 2, impacts the RS's responsiveness to external events, such as elections, where content curation must rapidly adjust to reflect evolving narratives. Algorithms constrained by inefficiencies default to existing patterns or rely on recommendations with a high success rate. To address these limitations, RS developers use adaptive temporal models, such as recurrent neural networks, which apply recursive methods to respond dynamically to changes in input (Xu et al., 2019; Wu et al., 2017; Devooght & Bersini, 2017; Beutel et al., 2018). However, in political discourse, this approach is inappropriate because narratives are shaped by cultural, linguistic, and sometimes ideological contexts specific to regions, which are arguably difficult to translate into algorithmic inputs. While adaptive solutions may adjust to temporal patterns, they often overlook nuanced political meanings or ethical considerations due to a lack of interpretability, which can only be effectively managed through consistent human intervention.

The further analysis suggests that, in the case of USB, RS outputs are associated with more insular linguistic patterns, which may coincide with reduced exposure diversity, without isolating algorithmic effects from content availability or platform practices. Across December to March, USB's average TTR is 0.058

and USR's is 0.052. While USR's higher TTR indicates slightly greater linguistic variety, both USB and USR maintain predominantly negative sentiment across periods. For example, USB's compound sentiment shifts from -0.425 in January to -0.374 in March, while USR remains consistently negative, with compound sentiment ranging from -0.373 in Dec+Jan to -0.195 in Feb+Mar. These trends suggest both discourses rely on similar engagement strategies, but USB's lower linguistic diversity could reinforce echo chambers more prominently. Algorithmic bias and overfitting, as discussed in Chapter 2, would further amplify these effects. Bias arises from RS reliance on historical user behaviour, favouring dominant narratives or high-engagement content. Overfitting may occur when recommendations are overly tailored to previously noted individual preferences, which do not exist in case of SP. Therefore, the sentiment analysis, when considered alongside the lower lexical diversity in USB, indicates that the RS produces a consistent pattern in Democratic party discourse that reduces exposure to diverse perspectives. Engagement loops amplify this issue by prioritising content aligned with prior user interactions, reinforcing patterns of linguistic similarity that may also align with the ideological consistency as discussed in Chapter 1. Mitigating these challenges often involves incorporating fairness-aware algorithms to enhance content diversity. While this aims to balance outcomes by prioritising specific perspectives over others, it introduces significant bias in contexts lacking sufficient nuance, which raises questions about whether distortions arise from the design of the RS itself or from external manipulation of the content it processes.

RS are associated with patterns of linguistic diversity, sentiment, and engagement observed in political content exposure. In the case of USB, the lower TTR observed across subgroups parallels patterns described in echo chamber research, where reduced lexical diversity can limit exposure to alternative perspectives. This diverges from much of the existing literature that links echo chambers to Republican-leaning contexts such as conservative talk radio or QAnon communities, which often employ simplified language and foster distrust of institutions (Jacobson, 2005; Hannah, 2021b, 2021a; Kayam, 2018; Benkler, Faris & Roberts, 2018; Jamieson & Cappella, 2008). By contrast, USB's slightly greater lexical variety in this dataset is associated with a wider range of vocabulary within recommended content. As discussed in Chapter 2, algorithmic bias and overfitting can amplify such effects by favouring high-engagement narratives and homogenising content. The metrics here indicate a less rigid approach to personalisation than might be expected, allowing for varied access to information. This is supported by the very low percentage of content recommended in the language of the country from which the SP's IP appeared to originate, suggesting that local language signals may have had minimal influence on recommendation trajectories. Such a distinction is consistent with the role of personalisation in shaping exposure to diverse perspectives while balancing user-specific recommendations with broader content diversity. Personalisation models that depend on extensive user data, as outlined in Chapter 2, have been shown to raise privacy and security concerns in RS research. Such a distinction is consistent with the role of personalisation in shaping exposure to diverse perspectives while balancing user-specific recommendations with broader content diversity. These dynamics also intersect

with the economic drivers of platform design, linking patterns of content curation to monetisation strategies and setting the stage for the following discussion of lexical density and its relationship to advertising incentives.

Economy

USB consistently demonstrates slightly higher lexical density than USR. This suggests that Democratic Party discourse prioritises denser, message-focused content, in contrast to Republican discourse, which shows broader lexical variety and a more distributed narrative structure. Such lexical density is indicative of optimised messaging and content design strategies, tailored to deliver concise and impactful communication within the discourse. The organised nature of USB's messaging within the recommendations coincides with patterns commonly associated with monetisable content, although this analysis cannot determine whether these features arise from RS optimisation or from upstream content production practices. This type of content is easier to target in an ad campaign, which in turn attracts more advertisers. RS optimisation processes may favour such density because it aligns with algorithmic efficiencies that reward high engagement rates. This case study explores how lexical density functions not only as a linguistic feature but also as a monetisation tool, enhancing the platform's profitability while shaping the user's political participation landscape. RS use predictive algorithms to match dense content with ad-driven engagement metrics, creating feedback loops that prioritise content with higher success rates (Steck, 2011). This is consistent with broader industry practices where economic incentives drive content curation strategies. While not directly evidenced in lexical density metrics, the pattern suggests an

implicit hypothesis that dense messaging is economically advantageous. For instance, a short and tightly framed political slogan can be readily repurposed across multiple recommendation streams and paired with targeted advertising, making it more cost-effective for both the platform and advertisers than longer, diffuse discourse. Such alignment may accidentally prioritise polarised or politically charged content due to its ability to sustain attention as discussed in Chapter 1, raising concerns about the broader impact on public discourse. Such alignment could influence the prominence of polarised or politically charged content due to its capacity to sustain attention, a factor explored further in the analysis of targeted engagement strategies. This dynamic is particularly evident when examining how sentiment patterns vary across audience groups, with certain demographics showing consistent exposure to a targeted messaging.

In this case study, male audiences engaging with USB content consistently encounter higher levels of negativity compared to their USR counterparts. This disparity indicates gender-associated differences in the sentiment of recommended content, which may reflect differential exposure patterns without establishing targeted optimisation or interaction effects. Negative messaging often produces stronger emotional responses, driving metrics such as time spent on content and click-through rates (Hansen et al., 2011). These approaches align with patterns in which male audiences are targeted as a profitable demographic for politically charged messaging, consistent with platform strategies to sustain user interest. Gender-specific targeting thus functions as a mechanism for optimising the relevance of content and advertisements, reinforcing engagement patterns shaped by both political

leanings and demographic characteristics. Demographic segmentation is a core feature of RS design, enabling platforms to align content delivery with advertiser demands (Zhang & Wedel, 2009). Gender-based targeting enhances ad relevance by tailoring content to specific audience segments, thereby increasing the likelihood of conversions. For male audiences, the combination of negative sentiment and politically charged discourse mirrors established advertising trends that favour emotionally engaging material. While a direct causal link between segmentation and ad revenue is not evidenced in the gathered data, the broader industry practice of using demographic data to drive revenue supports this hypothesis. By embedding demographic segmentation into RS workflows, platforms provide advertisers with precision targeting, reinforcing the economic rationale for gender-specific engagement. In practice, attempts to balance bias with demographic relevance through approaches such as inclusive targeting can introduce new distortions, as they may privilege some perspectives while constraining others. Metrics designed to balance diversity with engagement can reduce the reinforcement of polarising narratives, but while these solutions are effective in general advertising contexts, they may not be appropriate for political discourse. Political messaging often adopts marketing strategies aimed at maximising reach and persuasion, yet it carries distinct responsibilities linked to democratic principles and equitable representation. This dual function creates tension between engagement-driven optimisation and the broader civic role of political discourse.

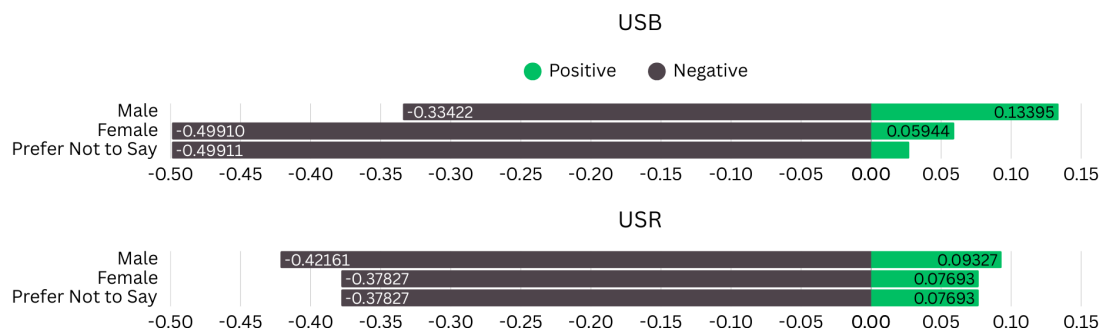


Figure 4a Average positive and negative sentiment scores by gender for USB and USR.

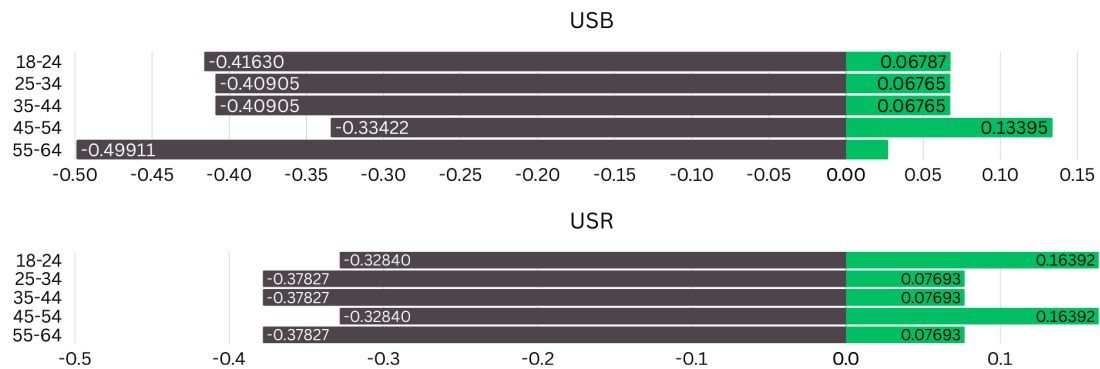


Figure 5b Average positive and negative sentiment scores by age group for USB and USR.

Based on observed linguistic patterns, the findings are consistent with USB exhibiting characteristics of an epistemic bubble, marked by linguistic homogeneity and persistent negative sentiment, where alternative perspectives appear limited rather than actively contested. This contrasts with USR, which displays broader lexical diversity and sentiment variation. The retention strategies identified in USB's corpus contribute to an environment where a limited set of narratives dominates, reflecting the platform's economic emphasis on high-engagement content. This may not reflect the full range of positions within the Democratic discourse, but rather the narrower subset of content that is most compatible with recommender optimisation, where emotionally charged and thematically consistent material is more readily amplified. RS are designed to maximise user retention, a critical metric for driving advertising revenue. Cost-sensitive algorithmic designs often favour such profitable content, justifying ad

pricing based on user activity metrics like time spent on the platform (Dave & Varma, 2014; Yuan, Wang & Zhao, 2013; Zhang, Yuan & Wang, 2014; Chen et al., 2011). This prioritisation of retention is associated with reduced content diversity, and the recommended content is dominated by polarising narratives. Such reinforcement of echo chambers is associated with constraints on exposure and epistemic agency, potentially shaping the conditions under which civic participation takes place, rather than constituting direct evidence of democratic harm. While engagement metrics serve immediate economic goals for the deployers of RS, their overuse is associated with reduced exposure to diverse perspectives, which may constrain the epistemic conditions relevant to pluralistic civic discourse. Although platforms are not explicitly designed to support democratic participation, political content moderation frameworks are often expected to uphold democratic values and prioritise inclusivity over profitability to preserve the integrity of public debate.

USB's consistent reliance on negative sentiment is consistent with repetition patterns commonly associated with profitability-oriented optimisation. These patterns suggest that USB's RS promotes repetition to potentially encourage predictable user behaviour. This aligns with advertising-based optimisation. In the context of political messaging, such repetition reinforces high-engagement narratives that increase platform use and generate revenue, without regard to political outcomes. In contrast, USR demonstrates broader lexical diversity and sentiment variation, indicating a strategy aimed at audience expansion. By diversifying content, USR appears to balance retention with outreach, catering to a wider audience base. These findings reflect the platform's

prioritisation of its own revenue goals, as it optimises for advertiser demands rather than fostering balanced discourse. The relationship between economic incentives and political content moderation is reflected in how profitability considerations shape RS design, linking platform revenue models to patterns in discourse curation.

The economic pressures shaping RS development are deeply intertwined with content curation strategies, as platforms optimise their designs to maximise profitability. Revenue-driven metrics such as user retention, click-through rates, and ad relevance dominate RS objectives, often at the expense of content diversity and balanced discourse (Abdollahpouri, Burke & Mobasher, 2019; Huang et al., 2020; Chen et al., 2021, 2023). Findings indicate a clear alignment between profit-driven strategies and content patterns. USB's reliance on negative sentiment and lexical repetition reflects engagement-driven profitability models, whereas USR's broader diversity points to a strategy focused on audience expansion. These approaches underline how economic incentives, compounded by technical debt, shape the linguistic and sentiment landscapes of political discourse. However, these strategies come with significant trade-offs. Prioritising profitability is associated with more homogeneous recommendation environments and reduced exposure diversity, without establishing effects on polarisation or democratic outcomes. The inherent tension between economic goals and democratic values presents a challenge for RS design. While contemporary systems prioritise short-term revenue gains, addressing the long-term need for inclusive, informed discourse remains largely unaccounted for. This tension is compounded by technical debt, which locks systems into existing

optimisation pathways. As a result, RS design tends to reinforce established biases rather than support transitions toward more diverse and equitable models of cultural participation.

Culture

Vocabulary size trends reveal notable differences in cultural inclusivity and representation across data. USB experienced a significant spike in vocabulary size in March, following a period of relative stagnation. This late increase coincides with a change in observed vocabulary size, without allowing inferences about intentional algorithmic adjustments or diversification objectives. In contrast, USR displayed consistent vocabulary growth over time, indicating steady exposure to a broader range of linguistic inputs. Such patterns suggest differing algorithmic strategies, with USB potentially responding to temporal shifts and USR maintaining a stable approach. Imbalanced datasets constrain the representation of minority cultural narratives, as training data often amplifies dominant perspectives while treating less frequent ones as noise to be filtered out. Such is consistent with USB's repeated linguistic patterns, which restrict exposure to diverse cultural representations despite a late increase in vocabulary size. Addressing data imbalances and fostering linguistic diversity requires intentional curation strategies. Periodic content diversification models have been shown in experimental RS research to increase exposure to underrepresented items and promote inclusivity (Nguyen et al., 2014; Abdollahpouri et al., 2019), but in political discourse such interventions risk compromising platform neutrality and masking systemic issues by introducing additional biases.

Middle-aged groups (categories 3 and 4) display the most pronounced partisan linguistic divides, with younger (categories 1 and 2) and older (category 5) groups showing greater linguistic similarity across partisan lines. This suggests that observed recommendation patterns vary across demographic categories, without establishing that algorithms are explicitly tailored to reinforce group-specific biases or preferences. These strategies are evident in the distinct content patterns for middle-aged groups, who are more likely to consume politically polarised content. The algorithmic focus on middle-aged users reflects their perceived value as a key demographic for political and commercial messaging. This optimisation aligns with the RS's objective to sustain engagement through targeted curation, which prioritises narratives likely to resonate strongly with this age group, at the expense of creating cross-generational discourse. Historic biases in training datasets further amplify generational divides. The datasets often reflect societal and cultural norms, embedding stereotypes that disproportionately affect middle-aged users. These biases promote narratives that align with historical content consumption patterns, which limit opportunities for intergenerational dialogue. Proxy bias amplifies these divides by associating demographic variables within the algorithm, such as age, with specific cultural narratives. Algorithms use these correlations to optimise recommendations that align with perceived generational preferences.

Similarly, sentiment trends in RS have been shown to shape the emotional tone of cultural narratives, as engagement-driven algorithms amplify affective content that sustains user attention (Hansen et al., 2011; Bakshy, Messing & Adamic, 2015). USB's higher negative sentiment with a range of fluctuations,

peaking in January, contrasts with USR's gradual consistency over time. These patterns reflect algorithmic feedback loops that amplify certain tones based on user engagement metrics. USB's persistent negativity may stem from critical narratives associated with discussion of possibility for Trump's second presidency. Meanwhile USR's sentiment moderation could reflect a strategy aimed at broadening appeal and promoting election participation. USB's volatility is associated with more variable emotional tones in recommended content, consistent with environments commonly described as echo-prone, without establishing reinforcement of group dynamics. In contrast, USR's more moderate sentiment may enable broader, less combative discourse. These dynamics suggest a potential role for RS in shaping emotional alignment among citizens, offering a plausible mechanism through which users' perceptions and interactions with cultural content could be affected. RS amplify specific emotional tones through engagement-driven recommendations. This amplification creates feedback loops where dominant sentiments, particularly negative ones, are reinforced. In USB's case, persistent negativity aligns with the platform's objective to maximise engagement, but at the cost of fostering polarised cultural narratives. Such reinforcement poses risks to inclusive cultural discourse in a democracy. This approach can restrict the diversity of cultural expression and embed emotional silos, which may reduce opportunities for shared understanding by producing a linguistic gap between discourse. This gap limits cross-group communication and reinforces fragmented public narratives. The long-term implication, in contexts where platforms such as YouTube function as major sites of political content exposure, concerns how algorithmically mediated environments may condition patterns of cultural expression, without establishing

effects on collective identity or democratic participation. To counteract sentiment polarisation, RS could integrate sentiment-balancing algorithms designed to diversify emotional tones in cultural content. However, applying such strategies to political discourse risks undermining neutrality and may deprive citizens of unmanipulated access to diverse perspectives.

Whereas USR shows gradual vocabulary growth and consistent sentiment patterns that indicate incremental broadening of its discourse, USB highlights the challenges of reactive and volatile algorithmic adjustments. These approaches reveal the tension between optimising user engagement and maintaining content diversity. This often leads to an amplification of dominant narratives while overlooking minority voices. Algorithms prioritise quantifiable metrics such as clicks and views, often at the expense of broader cultural representation. This dynamic creates feedback loops that amplify polarisation and limit cross-demographic discourse by creating a cultural and linguistic gap. Efforts to introduce balancing mechanisms, such as adaptive sampling or sentiment modulation, expose the risks of algorithmic intervention. In the context of democratic political discourse, this raises questions about whether a self-interested platform can ever act as a neutral intermediary.

Quality

Chapter 2 highlights the role of RS in influencing content diversity and user engagement through algorithmic design and prioritisation of metrics such as precision, diversity, and fairness. Sentiment patterns can be analysed through the lens of these broader themes. USB's predominant negativity reflects a

pattern where RS algorithms prioritise content designed to elicit strong emotional responses, aligning with the concept discussed in Chapter 2 that RS often optimise for engagement at the expense of diversity. USB's steadier sentiment trends indicate a less polarising approach, potentially attributable to algorithmic strategies aimed at maintaining stability. This aligns with the broader research discussion of trade-offs in RS design, which frame the challenge of balancing engagement-driven outputs with exposure diversity, while noting that platforms such as YouTube primarily optimise for engagement rather than for broader cultural outcomes (Cakmak, Agarwal & Poudel, 2025; Yu et al., 2024; Cakmak, Agarwal & Oni, 2024). Chapter 2's analysis of statistical discrimination can also help to interpret the sentiment disparities between USB and USB. By reinforcing dominant emotional tones, RS limit the range of sentiment exposure, marginalising less engaging or neutral narratives. This bias is particularly evident in USB's audience, where male demographics encounter more negative sentiment, illustrating how observed sentiment exposure differs across demographic categories, without establishing intentional demographic-specific targeting.

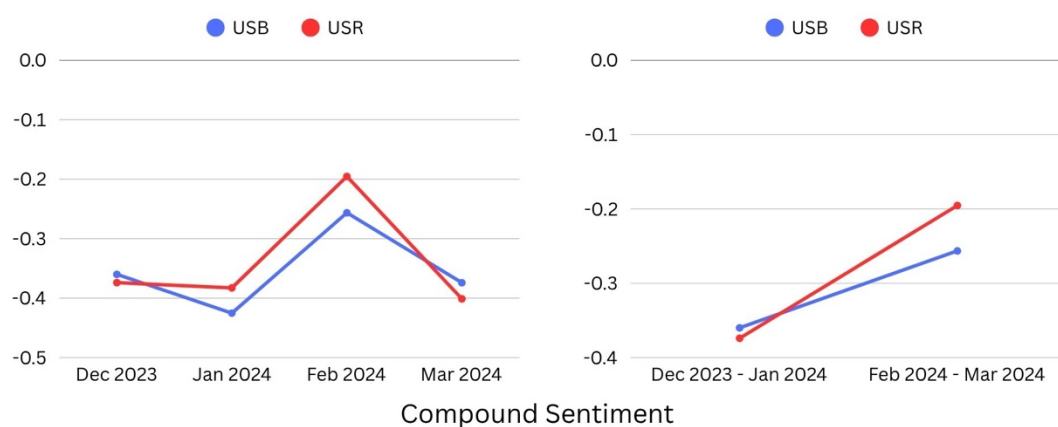


Figure 6 Compound sentiment scores for USB and USB.

Chapter 2 emphasises the role of RS in balancing user satisfaction and system efficiency through consistency in recommendations. USB's volatility in tone and narrative focus illustrates the risks associated with reactive algorithms. This unpredictability, shaped by sudden responses to external events or engagement spikes, challenges user trust and satisfaction. In contrast, USR's steadier trajectory reflects a more consistent pattern of discourse, which corresponds to the balance between accuracy and user expectations described in Chapter 2. The shrinkage effect, highlighted in Chapter 2 as a statistical approach to managing variability, plays an important role in achieving content consistency. By pulling recommendations toward an average, shrinkage minimises extremes, reducing volatility and creating a predictable user experience. For USR, this results in smoother sentiment trajectories and stable narratives that enhance user trust. However, USB's volatility suggests a limited application of shrinkage, leading to abrupt shifts that undermine consistency. While shrinkage promotes reliability, Chapter 2 warns against its overuse. For USB, the absence of mechanisms to balance shrinkage with diversity may amplify echo chambers, constraining epistemic agency. For example, USB's male group encounters consistently negative sentiment, reflecting dominant engagement patterns that shrinkage reinforces without introducing alternative perspectives.

Temporal patterns in USB and USR highlight the balance between immediate engagement and sustained user satisfaction, as discussed in Chapter 2. USB's sentiment volatility reflects a fast-decision-making approach prioritising transient trends during periods of high activity. In contrast, USR's gradual

sentiment progression aligns with long-term strategies designed to foster stability. These differences illustrate the trade-offs in temporal strategies that RS must navigate. As Chapter 2 outlines, fast decision-making algorithms optimise short-term user engagement by amplifying emotionally charged content. USB exemplifies this pattern, with observed recommendation outputs varying during peak periods, such as January, without allowing inferences about reactive algorithmic intent or strategic timing. While effective in the short term, this prioritisation risks neglecting broader narratives, creating a disjointed user experience. USR's steady progression reflects long-term strategies that align with Chapter 2's emphasis on reliability, likely shaped by differences in initial inputs and user signals rather than distinct platform optimisation goals. These variations can lead the same system to produce different outcomes across user groups. Meanwhile, the discussion of overfitting and underfitting provides a framework to understand temporal instability. USB's volatility is consistent with patterns commonly described as sensitivity to transient trends, where recommendations track short-lived content patterns, without establishing model-level overfitting or echo chamber intensification. This intensifies echo chambers, narrowing narrative diversity. On the other hand, underfitting, where generalisation leads to a lack of nuanced recommendations, can result in irrelevant content. USR's steadiness indicates a more balanced approach, mitigating these risks by maintaining relevance across temporal shifts.

Chapter 2 highlights how biases in RS evaluation metrics, such as click-through rates and time-on-platform, reinforce narrow content loops. USB's reduced linguistic variety illustrates the risks of metrics optimised for

engagement, where algorithms favour high-performing narratives at the expense of underrepresented content. This dynamic perpetuates systemic inequalities, constraining thematic diversity and marginalising less dominant cultural perspectives. USR's broader range suggests a more balanced approach, but as discussed in Chapter 2, achieving true inclusivity requires fundamental shifts in RS evaluation frameworks. Metrics incorporating novelty, while promising, must overcome existing biases embedded in engagement-based evaluation methods.

The quality metrics used in RS profoundly influence political and cultural discourse by shaping user engagement and narrative exposure. Metrics such as accuracy, engagement, and consistency play a pivotal role in determining the content users encounter, reinforcing thematic and emotional patterns. Although these metrics enhance measurable aspects of user satisfaction and platform efficiency, they also contribute to challenges such as echo chambers, sentiment imbalance, and limited diversity, particularly in politically charged environments. Balancing stability, diversity, and sentiment presents a significant challenge for RS design. Stability fosters user trust but risks stagnation without periodic injections of diverse content. These tensions highlight the need for RS to evolve beyond conventional quality measures toward frameworks that incorporate broader cultural and ethical considerations.

The US case study suggests contrasting outcomes in how the RS shapes political discourse on YouTube. These differences likely emerge not only from the optimisation processes of the algorithm but also from the nature of the input content, where Democratic-leaning material in USB appears more repetitive and

sentimentally charged, while Republican-leaning material in USB presents greater lexical variety and steadier tone. USB demonstrates lower lexical diversity and persistent negativity, reinforcing echo chambers and limiting exposure to diverse perspectives, while USR maintains higher diversity and gradual sentiment moderation, fostering broader engagement. These differences underscore how RS-driven content curation prioritises engagement over inclusivity by creating a linguistic discourse gap, which is amplifying polarisation in USB while offering a more varied narrative experience in USR.

Research on political communication consistently links Democratic discourse to education, complexity, and diversity, while Republican discourse is often described as narrower, repetitive, and linguistically constrained (Sclafani, 2017; Kayam, 2018; Zhou et al., 2024; Ahmadian, Azarshahi & Paulhus, 2017; Terren & Borge, 2021; Justwan et al., 2018; Barberá et al., 2015; Levendusky, 2009; Parker, 2019; Nie, Junn & Stehlik-Barry, 1996; Glaeser, Ponzetto & Shleifer, 2007; Grossmann & Hopkins, 2015; DeBray, 2006; Jiang et al., 2024; Wojcieszak & Mutz, 2009; Sylwester & Purver, 2015; Stromer-Galley, 2003; Waytz et al., 2019; Jost, 2006; Demszky et al., 2019). A broad body of studies shows that Democratic language tends toward greater cognitive complexity and linguistic variety, reflecting both the sociocultural composition of the Democratic base and inclusive messaging strategies (Conway et al., 2016). Ahmadian, Azarshahi and Paulhus similarly demonstrate that liberal discourse is more abstract and analytical, while conservative discourse relies on concreteness and repetition (Ahmadian, Azarshahi & Paulhus, 2017). These findings have informed expectations that USB may exhibit higher lexical diversity and more expansive

narratives, while USR may display more repetitive and affect-driven styles, without treating these patterns as inherent or deterministic. Analyses of populist rhetoric, including Trump's speeches, reinforce this view by highlighting low readability levels and direct phrasing that resonate with audiences but appear linguistically limited. The present dataset does not overturn these accounts but shows how RS alter communicative outcomes. USB discourse, while grounded in the educated and diverse tradition described above, becomes locked into repetitive cycles of commentary on events that already have established interpretations within the Democratic community (Demszky et al., 2019). Studies of partisan responses to mass shootings, for example, show that Democratic discourse repeatedly emphasises frames such as gun control and victim-centred narratives (Demszky et al., 2019; Zhang et al., 2023). Offline, such consistency reflects consensus and authority. Online, however, RS amplify repetition, narrowing the range of expression and reducing lexical diversity. What functions as discursive consistency in traditional media is associated, in recommendation environments, with more repetitive exposure patterns, without establishing a causal transformation by RS. The contrast with USR is notable. Trump's communication style, often described as simplistic, is also marked by spontaneity, shifting phrasing, and anecdotal variation. Ahmadian, Azarshahi and Paulhus show that conservative discourse draws on concrete, image-rich language that can introduce diversity despite lower cognitive complexity (Ahmadian, Azarshahi & Paulhus, 2017). Within RS, this variability generates broader lexical diversity, as recommendations draw from a wider linguistic pool. Consequently, USR content in this dataset appears more varied than USB, reversing sociolinguistic expectations. This asymmetry extends to cross-group

exposure. The data show that USR audiences were frequently recommended USB content, while USB audiences were rarely exposed to USR material, a pattern consistent with studies documenting uneven exposure to opposing views online and its role in reinforcing polarisation (Bail et al., 2018). The findings therefore suggest that RS can invert sociocultural patterns: liberal discourse associated with diversity offline can appear narrow online, while conservative discourse associated with repetition offline can generate broader variety in algorithmic curation.

3.3 Russia Case Study

During the case study period, Russia's political landscape was influenced by the ongoing war in Ukraine and the 2024 presidential elections, held from 15 to 17 March. President Vladimir Putin faced no significant opposition, as alternative candidates belonged to the 'системная оппозиция' [system opposition], state-controlled political parties designed to simulate democratic competition. The domestic narrative, shaped by state-controlled media, portrayed Russia as a defender against Western aggression, a protector of the Russian world (Russian culture), and an 'upholder of 'традиционные ценности' [traditional values] in a divided world.' Despite international sanctions and the withdrawal of foreign businesses, Russia's economy demonstrated resilience (Prokopenko, 2024). Many citizens appeared to accept economic challenges, framing them as a necessary cost in what was described as a 'conflict with NATO' (Melkozorova, 2024). This narrative was central to the political discourse leading up to the elections, coinciding with increased propaganda and the prominence of

algorithmically curated content, without establishing effects on public opinion. This study focuses on two video-hosting platforms, YouTube (RUB) and Dzen (RUR). Dzen became Russia's primary platform after integrating with VK's video archive (the biggest social network in Russia), offering highly structured recommendations. In contrast, YouTube's Russian segment faced government censorship but remained accessible to anti-war content, making it the only platform where such content could be found. Pro-war narratives were exclusively analysed on Dzen, while anti-war narratives were sourced from YouTube. Similar to the US case study, both RUB and RUR discourses were analysed for linguistic diversity, sentiment patterns, and demographic-specific content. This case study illustrates how similar recommender technologies, operating within distinct political and socio-cultural environments, are associated with different patterns of linguistic exposure in politically significant discourse.

Technology

Across December to March, RUR's average TTR is 0.1654 and RUB's is 0.1179, with RUR higher in December, January, and markedly in February, and slightly higher in March. Vocabulary size also varies by month, with RUB exceeding RUR in December, January, and March, while in February RUR is higher. In March, the sizes are closer: RUB 72,930 and RUR 63,337. This similarity in size coincides with a difference in TTR: RUB 0.0961 and RUR 0.1043.

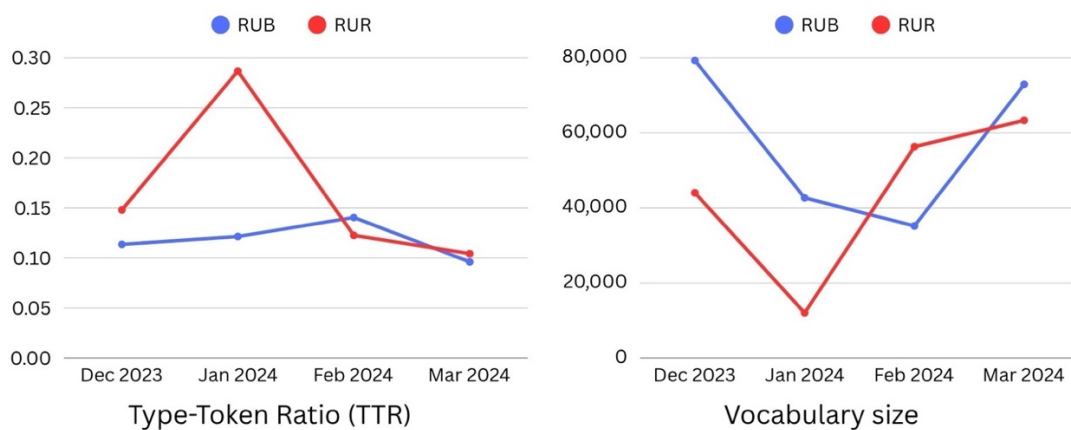


Figure 7 TTR and vocabulary size for RUB and RUR.

The temporal aggregation, therefore, suggests that RUR content was lexically more varied than RUB's for most of this period. When TTR is recalculated for demographic subgroups, RUR retains this advantage across all age categories, though the margin is smaller than monthly averages suggest. This indicates that the higher aggregate TTR for RUR is consistent across groups rather than being driven by isolated demographic spikes. The data therefore descriptively indicates a pattern in which RUR exhibits higher observed lexical variety than RUB, both overall and across demographic segments, without implying statistical robustness. As in the USB–USR case, this data presents an example of the evaluation metric fallacy: a global metric suggests a strong advantage for RUR in lexical diversity, but subgroup analysis reveals a more nuanced distribution. Optimisation for aggregate performance can conceal distributional imbalances, meaning that while the platform may appear to provide broader linguistic variety overall, many demographic segments experience a narrower range of lexical input. As explained in the US case study, occasional high-diversity events inflate overall averages, while subgroup streams recycle more limited language. The

same mechanism is evident here, with RUR's aggregate advantage masking constrained lexical input across its demographic subgroups.

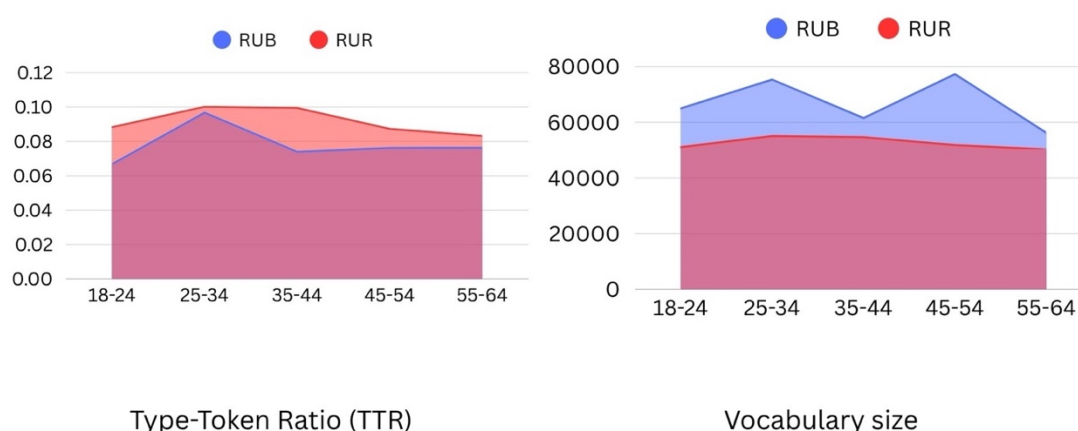


Figure 8 TTR and vocabulary size by age groups in Russia Case Study.

The temporal metrics of RUB and RUR reveal distinct differences in content evolution and sentiment dynamics. RUR exhibits significant temporal fluctuations in lexical diversity and vocabulary size, suggesting algorithmic flexibility. This variability reflects Dzen's capacity to adapt its recommendations dynamically, allowing for greater narrative diversity and responsiveness to external events. By broadening the scope of content over time, RUR fosters a more varied user experience, highlighting the platform's ability to engage with evolving discourse. In contrast, RUB shows steadier lexical diversity but greater volatility in sentiment, marked by sharp shifts between months. This pattern indicates YouTube's focus on optimising engagement metrics through emotionally charged content cycles. While this approach may stabilise certain engagement loops, it also amplifies sudden sentiment swings, limiting the platform's ability to maintain consistent tonal balance. Such patterns are consistent with recommendation environments often described as echo-prone, without establishing reinforcement effects or reductions in content exploration.

RUR demonstrates greater linguistic diversity, reflecting a less restrictive approach to content selection. By enabling variation in lexical range, RUR supports a wider set of narratives, potentially facilitating user interaction across more perspectives. This pattern suggests some flexibility in balancing engagement with content variety, although the narratives remain aligned with a pro-war framing. In contrast, RUB maintains comparable lexical diversity month to month but shows greater volatility in sentiment, with sharper shifts between positive and negative tones. This sentiment pattern, combined with more polarised recommendation outputs, is associated with narrower exposure conditions often described as echo-prone, without establishing effects on user behaviour or discourse breadth.

RUR's personalisation patterns result in broader lexical diversity and a wider thematic scope than RUB, allowing users to encounter a greater variety of narratives. This breadth, however, remains framed within a consistent pro-war perspective, limiting the diversity of interpretive angles despite topical variation. RUB's curation is more stable in lexical diversity but prioritises engagement patterns that coincide with sharper sentiment shifts and reinforce oppositional, anti-war narratives. This focus narrows thematic scope and concentrates attention on critique, maintaining a polarised frame. These differences illustrate how platform-specific personalisation strategies shape both the range and framing of content, influencing the balance between user exposure to varied perspectives and the reinforcement of existing viewpoints.

The RS on platforms such as Dzen and YouTube shape both user experiences and the circulation of political narratives. RUR's higher lexical diversity and marked temporal shifts in type–token ratio indicate an algorithmic configuration that can broaden narrative exposure, although this occurs within a consistently pro-war frame. RUB's steadier lexical diversity and sharper sentiment swings suggest a more constrained curation that reinforces polarised and repetitive patterns, with an emphasis on oppositional narratives. These contrasts show how algorithmic design affects lexical variety, temporal dynamics, and the framing of content. Balancing engagement, diversity, and personalisation remains a structural challenge. RUR's lexical variation occurs within a consistently pro-war orientation, while RUB's more constrained curation operates within an anti-war frame. The relationship between political orientation and observed linguistic diversity is correlational within this analysis, as the study cannot isolate whether recommendation processes drive stance reinforcement or merely reflect existing content distributions. Both cases illustrate trade-offs in RS design, where gains in one objective may reduce performance on another.

Economy

RUR consistently exhibits higher lexical density, meaning a greater proportion of content words relative to function words. While this does not necessarily equate to higher informational density, in this case it reflects content designed to convey more concentrated messaging within shorter segments. This design appears optimised for retention by presenting concise, content-rich messaging that sustains engagement within shorter sessions. The platform's focus on delivering dense, information-heavy narratives may reflect an attempt

to attract and retain users over shorter interaction periods by maximising engagement within each session. In contrast, RUB demonstrates lower lexical density, often shaped by the dominance of long-format content such as live broadcasts and extended commentary streams. These longer sessions rely on repetitive phrasing and extended narrative loops, which reduce lexical variety relative to length. RUB's lower lexical density reflects the nature of its recommended content, which consists of extended, less dense loops that support longer viewing sessions. While effective for sustaining broad user engagement, it relies on repetitive patterns that limit informational value and thematic variety. These contrasting approaches illustrate how platforms adjust their content strategies based on user interaction patterns and engagement goals. Algorithms on both platforms optimise content density to maximise ad placements and click-through rates. RUB's lower density approach enables longer engagement windows, creating more opportunities for ad placements across extended viewing periods. This strategy aligns with YouTube's broader monetisation goals, favouring continuous user interaction over dense, high-impact messaging. In contrast, RUR's high lexical density supports its strategy of targeting users likely to engage with dense, information-driven narratives, aligning well with the objectives to disseminate the information on the government controlled and curated platform. By focusing on shorter, high-retention sessions, Dzen produces recommendation outputs that concentrate politically aligned content within limited timeframes, without establishing intent or effects on viewer perception. Both platforms demonstrate how lexical density is calibrated to serve the dual objectives of user engagement and revenue generation, though at the expense of content diversity.

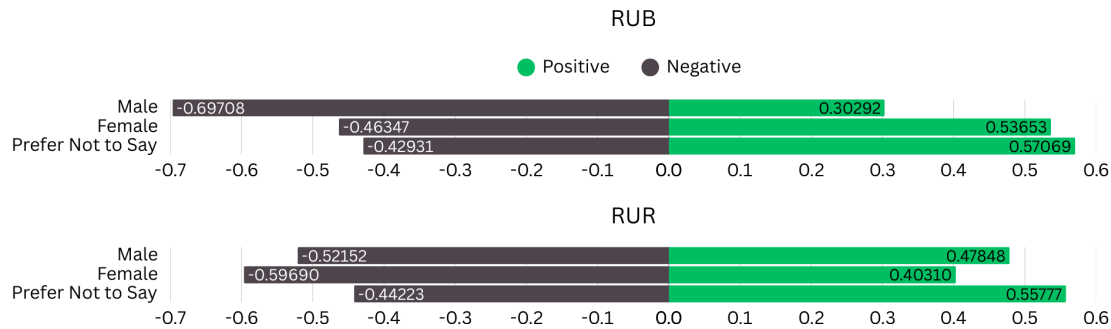


Figure 9a Average positive and negative sentiment scores by gender for RUB and RUR.

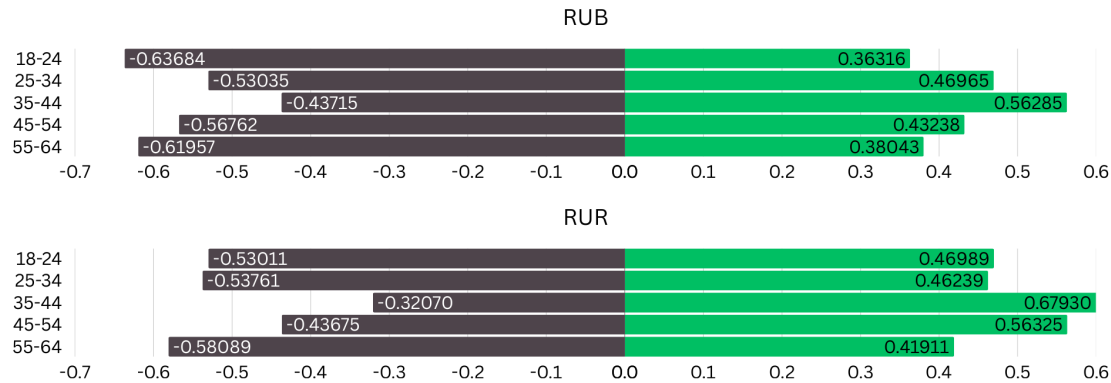


Figure 10b Average positive and negative sentiment scores by age group for RUB and RUR.

RUR exhibits stronger gender-targeted content, particularly in male and neutral categories, aligning closely with advertiser demands for high interaction metrics. Observed recommendation patterns on RUR differ across gender categories and coincide with higher engagement-oriented metrics, without establishing intentional tailoring to gender-specific preferences. Demographic segmentation shows that younger age groups on RUR encounter higher sentiment variability and lexical diversity, reflecting a tailored strategy for engaging with dynamic narratives. In contrast, RUB recommendations display greater polarisation among middle-aged users, with consistent negative sentiment reinforcing entrenched ideological divides. Notably, RUR did not recommend any female-targeted content throughout the study period. By contrast, in all three other cases female accounts were consistently steered toward gendered content such as cartoons and prescriptive material on women's

roles. This gap reflects an evident alignment with the Russian emphasis on war, military-related topics, and the demographic groups that such content predominantly targeted for (Meduza, 2024; Cottiero et al., 2015; Melkozorova, 2024; Prokopenko, 2024). This approach demonstrates the platform's focus on high-value user segments that generate consistent interaction. In contrast, RUB employs more generalised demographic strategies, reflecting its broader user base and emphasis on diverse engagement. By focusing less on segmentation, RUB accommodates a wider audience but sacrifices the precision of targeting seen on RUR. These differences in gender-based targeting highlight the varying monetisation strategies between the platforms, with RUR prioritising high-intensity interaction and RUB focusing on overall reach.

RUR is characterised by a pro-war narrative that exhibits high levels of emotional reinforcement and polarisation. Rather than indicating a shared retention strategy across platforms, this pattern reflects how polarised content aligns with engagement metrics within Dzen's state-aligned and commercially integrated ecosystem. Such content sustains repeated exposure and interaction, which is compatible with advertiser-oriented incentives, without implying that polarisation itself is the sole or uniform mechanism driving retention. RUB, while exhibiting an opposing ideological orientation, also displays polarised linguistic patterns, though these emerge within a different structural context. In this case, polarisation reflects the aggregation of oppositional content produced by independent creators rather than platform-level optimisation for retention. Engagement loops are present, but they arise from content dynamics and audience alignment rather than from targeted monetisation strategies. The

comparison therefore highlights that polarisation can be present across competing platforms while serving distinct functions shaped by their respective regulatory, economic, and governance environments.

Economic incentives shape recommendation environments by privileging content that sustains engagement and repeated exposure, which can indirectly reinforce polarisation and repetition. In both platforms, engagement-oriented optimisation coincides with reduced exposure to divergent perspectives, though the mechanisms through which this occurs differ. In the RUB case, oppositional content circulates within commercially mediated recommendation loops that may narrow discourse despite the ideological diversity of producers. These patterns illustrate how economic optimisation can constrain exposure even in contexts where content is not centrally coordinated, highlighting the tension between engagement-driven curation and the maintenance of discursive breadth.

Culture

RUR demonstrates steady growth in vocabulary size, peaking in March. This indicates broader topic coverage and inclusivity, despite operating with a smaller corpus compared to RUB. The increase in vocabulary suggests a dynamic approach to content curation that expands thematic breadth over time. This pattern supports linguistic diversity, enabling users to engage with a wider range of narratives. RUB data shows repetitive, state-centred phrases across all demographics, such as ‘российская федерация’ [Russian Federation], ‘европейский союз’ [European Union], and even duplications such as ‘украина украина’ [Ukraine Ukraine]. The variation across age and gender is minimal, and

some lexical items likely originate from automated transcripts or repeated boilerplate phrases rather than genuine narrative variation. In contrast, RUR presents a more varied set of military and technical terms, including ‘артиллерийская установка’ [artillery system], ‘самоходный артиллерийский’ [self-propelled artillery], and ‘беспилотный летательный аппарат’ [unmanned aerial vehicle (drone)], which reflect situational updates in wartime communication. This difference reflects a cultural approach to discourse, with RUB reinforcing stable national narratives while RUR adapts to situational updates in wartime communication. Despite drawing on a broader dataset, RUB exhibits limited linguistic diversity, suggesting that its algorithms favour established content loops. This strategy prioritises engagement metrics but restricts exposure to underrepresented cultural and linguistic themes. As a result, RUB is associated with more homogeneous linguistic patterns that coincide with a narrower range of expressed perspectives, without establishing effects on inclusivity or polarisation. Imbalanced datasets significantly affect cultural diversity in RS outputs by favouring dominant narratives. This bias is evident in RUB’s static vocabulary, which prioritises widely consumed content while marginalising minority perspectives. Such imbalances create a feedback loop where the most popular themes receive disproportionate amplification, narrowing the platform’s cultural inclusivity, which in particular makes this crucial information unreachable for the opposing side. RUR, despite its smaller corpus, compensates with richer linguistic variety. The platform’s algorithms appear to prioritise diverse thematic representation, enabling broader cultural participation.

Younger users on RUR encounter higher linguistic diversity compared to RUB, reflecting platform-specific strategies tailored to engage younger demographics. RUR's dynamic approach to content curation introduces varied narratives, fostering broader engagement among this age group. This diversity enhances cultural participation, aligning with RUR's emphasis on promoting thematic variety despite its smaller corpus size. In contrast, middle-aged users on both platforms display stronger alignment with polarised content. Recommendation outputs on both RUR and RUB differ for this demographic and are more frequently associated with high-engagement content, without establishing intentional prioritisation or effects on exposure diversity. This targeted approach underscores the platforms' focus on retaining middle-aged users through emotionally charged and ideologically consistent narratives. As discussed in the US case study, historic biases embedded in training datasets perpetuate generational divides by reinforcing traditional consumption patterns and limiting intergenerational engagement. RUR exhibits higher sentiment variability, reflecting a dynamic discourse shaped by external events. This variability allows the platform to engage users with diverse emotional tones, broadening their exposure to a range of narratives. The shifting sentiment patterns align with RUR's approach to fostering dynamic engagement through topical and emotionally varied content. In contrast, RUB demonstrates sentiment patterns that fluctuate between mildly negative and moderately positive, with monthly compound scores ranging from -0.18 to 0.29. While this variation can maintain user engagement, it reduces tonal consistency and may limit the platform's capacity to sustain a balanced emotional range in its cultural

narratives. By prioritising consistency, RUB reinforces homogenised discourse, reducing the emotional richness of its cultural narratives.

The cultural implications of RS design on platforms like Dzen and YouTube highlight their significant impact on linguistic diversity, demographic engagement, and narrative evolution in politically charged contexts. RUR's dynamic approach fosters broader narrative adaptability and greater linguistic diversity, enabling engagement with evolving cultural discourse. However, this flexibility is associated with the circulation of more polarised ideological content, particularly on a government-owned platform where political messaging is already tightly controlled, without establishing amplification effects attributable to RS. In contrast, RUB's stricter algorithmic constraints reinforce static narratives and dominant themes, limiting inclusivity and responsiveness to external changes. The focus on consistency in RUB demonstrates how engagement-driven metrics can restrict diversity, creating cultural silos that undermine collective identity. On Youtube, anti-war discourse remains formally available but is difficult to encounter without prior alignment, since algorithms prioritise content that maximises engagement and therefore amplify dominant narratives. As a result, anti-war material circulates mostly within audiences already opposed to the war, rather than reaching broader publics. This confinement is associated with more closed circulation patterns, where oppositional content primarily reaches already aligned audiences, without establishing effects on civic influence or regime contestation.

Quality

While RUR demonstrates a more dynamic and adaptable approach, RUB prioritises stability and predictability, creating trade-offs in user engagement and cultural participation. Sentiment analysis reveals distinct algorithmic strategies between RUR and RUB. RUR exhibits significant sentiment variability, with positivity peaking in January (0.49) and declining to 0.43 in March, while negativity rises correspondingly. This variability reflects a dynamic approach to content adaptation, as noted in Chapter 2's discussion on novelty metrics. By engaging with diverse emotional tones, RUR broadens the spectrum of user experiences, fostering dynamic engagement through topical shifts. In contrast, RUB's sentiment profile is more variable, with periods of predominantly negative tone interspersed with occasional positive peaks. This fluctuation sustains engagement and is associated with more polarised emotional framing in recommended content, without establishing effects on discourse breadth or cultural participation. Chapter 2 highlights the risks of prioritising engagement metrics at the expense of emotional diversity. RUB's stability ensures predictable recommendations but reinforces polarised emotional environments, narrowing the breadth of discourse and restricting cultural participation. This aligns with Chapter 2's caution about echo chambers, where algorithmic amplification of specific emotional tones limits exposure to alternative perspectives. The contrast underscores how sentiment management reflects broader algorithmic priorities. While RUR leverages variability to adapt to shifting discourse, RUB's rigidity risks homogenising emotional narratives, reducing the richness of its cultural outputs.

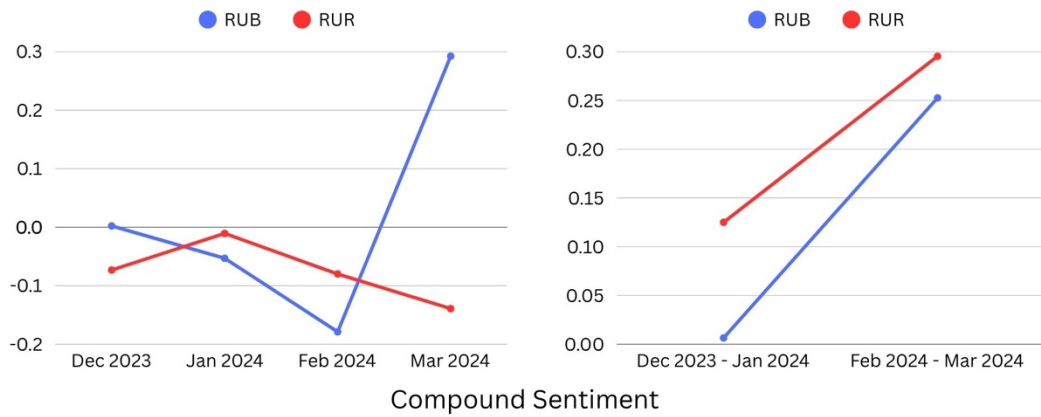


Figure 11 Compound sentiment scores for RUB and RUR.

Chapter 2 emphasises the importance of reliability in RS, where consistent recommendations build trust and ensure user satisfaction. RUR achieves this balance by delivering steady content despite sentiment shifts. Its vocabulary growth from 12,038 in January to 63,337 in March illustrates its adaptability while maintaining thematic variety. This approach aligns with Chapter 2's focus on balancing accuracy and novelty, where consistency is coupled with diversity to sustain engagement without alienating users. In turn, RUB exhibits more stable thematic patterns associated with its anti-war focus. This is consistent with concepts discussed in Chapter 2 regarding sensitivity to narrow content distributions, without establishing model-level overfitting. While RUB's stability ensures predictable user experiences, it sacrifices the dynamism needed for broader engagement. Such rigidity reinforces Chapter 2's warning about stagnation risks in RS, where over-reliance on stability limits thematic exploration. The distinction between RUR and RUB reflects the trade-offs in content consistency. RUR's balanced approach fosters adaptability alongside trust, while RUB's rigidity prioritises retention over dynamic engagement, constraining its ability to accommodate broader discourse.

Temporal stability, as outlined in Chapter 2, involves balancing responsiveness to external influences with cohesive content delivery. RUR demonstrates significant temporal adaptability. Its expanding vocabulary and dynamic sentiment shifts are associated with changing recommendation outputs over time, without establishing intentional alignment with user preferences or cultural trends. This adaptability highlights Chapter 2's discussion on novelty metrics, where responsiveness enhances engagement by reflecting ongoing societal discourse. RUB, in contrast, exhibits temporal rigidity, maintaining consistent but narrow content loops. This static approach prioritises stable engagement metrics, limiting its ability to reflect cultural shifts. Chapter 2's exploration of underfitting captures this challenge, where overly generalised algorithms fail to accommodate nuanced changes in discourse, leading to disengagement. The divergence between RUR and RUB illustrates the importance of temporal flexibility. RUR's ability to evolve with cultural narratives fosters engagement and relevance, while RUB's rigidity risks alienating users seeking dynamic and varied content.

Sentiment variability, content consistency, temporal adaptability, and thematic breadth all contribute to how RS influence cultural discourse. RUR demonstrates dynamic adaptability in sentiment and thematic range, fostering broader engagement and reflecting evolving user interests. In turn, RUB maintains a more rigid structure, prioritising stability and thematic consistency but limiting opportunities for narrative diversity and innovation. Balancing engagement with diversity and stability remains a key challenge, particularly in

politically charged environments. While RUR's flexibility supports dynamic content delivery, it risks amplifying polarised narratives. RUB's focus on consistency reinforces predictable user interaction but at the cost of thematic and emotional breadth. These tensions underscore the difficulty of designing RS that effectively balance user retention with inclusive and evolving discourse.

The Russian case study highlights how RS shape political discourse across two platforms, RUR (Dzen) and RUB (YouTube). RUR, integrated with VK's video archive, promotes pro-war narratives through algorithmically driven content that displays broader lexical diversity and sentiment variability, reflecting its adaptability to shifting wartime events. RUB, by contrast, provides a space for anti-war discourse where content is accessible but circulates in stable, repetitive loops that prioritise consistency over diversity. What makes these findings significant is that they replicate the same dynamics identified in the US case: RS behave in strikingly similar ways across democratic and authoritarian contexts. In both cases, the apparent sociocultural assumptions are inverted. Just as Republican discourse, often characterised as narrower and more repetitive, showed broader lexical diversity than Democratic discourse on YouTube, so too RUR, commonly perceived as less educated than RUB, produced more varied linguistic outputs. These parallels underscore that the mechanisms of RS, rather than national political culture alone, drive the reversals observed. The comparison highlights how similar recommendation processes can be associated with output patterns that differ from offline expectations across contexts, without establishing effects on polarisation or inclusivity.

3.4 Conclusion

This chapter examined how RS shape the conditions of user engagement, linguistic diversity, and opportunities for civic participation across technological, economic, cultural, and qualitative dimensions. Analysing YouTube and Dzen, it showed how these systems configure the information environments users encounter, shaping exposure to particular viewpoints, emotional tones, and cultural narratives. The findings indicate that engagement-optimised systems can produce outcomes that privilege certain perspectives while limiting others, with economic imperatives often shaping these trajectories. In the Russian case, Dzen's pro-war narrative aligned with state objectives, and its monetisation model depended on maintaining this alignment, since advertiser support in an authoritarian context is contingent on the platform's consistency with government-approved narratives. On YouTube, both USB and USR content displayed emotionally charged narratives, consistent with advertising-driven optimisation strategies that favoured engagement metrics over sustained diversity. These patterns illustrate how revenue-focused curation can narrow informational variety, creating conditions in which balanced discourse is less visible. In both platforms, the cold start problem was a notable driver of early content homogeneity, as new accounts were directed towards highly polarised recommendations in the absence of prior activity signals. These observations describe differences in recommendation outputs and exposure patterns rather than demonstrating adaptive system behaviour or isolating recommender effects from changes in the underlying content supply. This contributed to a widening civic empowerment gap, with the proxy accounts in this study encountering

narrative reinforcement rather than substantive diversity. The reliance on historical patterns in training data further reinforced these trajectories, embedding existing biases into recommendation outputs and constraining access to a broader range of perspectives. The US case study considered how YouTube's RS shaped political discourse during the 2024 presidential primaries, focusing on the interaction between partisan segmentation, linguistic diversity, and sentiment patterns. The Russian case study examined how RS influenced political narratives during the 2024 presidential election, highlighting the interplay between demographic targeting and ideological framing. In the US, younger users encountered a somewhat broader range of content, whereas middle-aged users in both USB and USR were more frequently associated with polarised recommendation outputs, without establishing targeted reinforcement of established positions. In Russia, younger RUR users engaged with a greater range of vocabulary and sentiment variation, although this diversity was embedded within a consistent pro-war frame. Middle-aged users in RUR and RUB experienced narrower thematic variety and more stable ideological positioning, consistent with retention strategies favouring repetition. Gender-specific patterns also differed. In the US, male users in USB more frequently engaged with negative and emotionally charged content, a pattern aligned with engagement-driven advertising models. In Russia, gender tailoring was less pronounced. RUR omitted female-specific targeting entirely, reflecting its emphasis on narratives positioned towards male audiences in military and patriotic contexts. Across both settings, gender-based curation aligned with platform retention goals but also reinforced pre-existing audience biases. Civic participation emerged differently across the two contexts. In the US, USB's

recommendations tended to align closely with partisan boundaries, reducing the likelihood of cross-party content exposure and limiting opportunities to engage with alternative perspectives. USR presented a somewhat broader lexical range but still prioritised retention through emotionally engaging narratives, which constrained the depth and nuance of political discussion. In Russia, RUB's focus on anti-war narratives restricted their reach largely to ideologically aligned audiences, limiting the circulation of oppositional viewpoints. RUR, in contrast, presented pro-war messaging that was dynamically adapted for engagement while remaining aligned with state priorities. In both cases, RS outputs revealed asymmetries in how political positions were represented, with system configurations that favoured polarisation over the facilitation of balanced civic dialogue. Cultural narratives were shaped through sentiment reinforcement, with RUR maintaining positive sentiment around pro-war themes to strengthen state-framed cohesion, and RUB adopting a predominantly negative tone in its critique of such narratives. USB's emphasis on emotionally charged content contributed to divisive discourse, whereas USR's steadier sentiment trends supported a more sustained, if still selective, narrative engagement. These patterns indicate that RS can embed sentiment strategies into content delivery, shaping cultural participation in ways that consolidate polarised identities and limit spaces for inclusive debate.

The two case studies indicate that RS can treat opposing positions differently even when addressing the same topic, with outcomes shaped by platform-specific configurations. Despite these variations, the underlying technical objectives of optimising for engagement and retention remain

consistent and produce broadly similar structural effects. This consistency suggests that system behaviour is shaped more by technological constraints than by the substantive content of recommendations. Shared training practices and optimisation methods, as outlined in Chapter 2, contribute to the amplification of echo chambers and the recurrence of tightly framed content loops. These patterns point to architectural limitations in balancing precision with diversity, particularly in politically sensitive contexts. In USB and RUB, reducing repetition, increasing lexical variety, and diversifying narrative framing could broaden exposure beyond ideologically aligned audiences. While RS do not act directly on speech production, their ranking and selection processes can influence which linguistic features gain visibility. Adjustments that favour greater variation in vocabulary and framing may help to reduce reinforcement cycles and expand access to alternative perspectives. Such changes could support more inclusive participation by enabling discourse to reach across demographic and ideological boundaries. The findings suggest that current RS development methods are not optimally configured for politically sensitive material. Designing systems intended to operate in politically sensitive contexts may require clearer articulation of curation rules and explicit consideration of how diversity and accessibility are operationalised. This sets the stage for Chapter 4, which examines policy approaches to AI, with a focus on measures aimed at addressing bias and fostering equitable outcomes in systems that affect democratic processes.

Chapter 4. Recommender Systems: AI Policy Frameworks

Previous chapters analysed both the technical challenges in RS development and the observable effects of their outputs when content was localised through synthetic profiles set to the USA and Russia. These platforms function as significant channels for citizens' cultural participation, which is relevant to analyses of civic engagement in contemporary media environments. The cultural impact of RS has been explored in research literature, with particular attention to their role in shaping access to and diversity of information (Bozdag & Van Den Hoven, 2015). Analysing how different jurisdictions regulate such systems is therefore critical to understanding their influence on democratic processes and the ways in which regulatory priorities reflect, reinforce, or challenge existing cultures of civic participation. This perspective makes it possible to examine how regulatory approaches respond to RS-related shifts in public discourse and how they may contribute to shaping the scope of civic engagement that these systems support.

This chapter undertakes a comparative analysis of regulatory approaches to AI systems in jurisdictions with differing political, legal, and cultural contexts, focusing on how these frameworks frame epistemic and civic risks associated with RS and how they affect epistemic agency. Building on the development challenges discussed in Chapter 2 and the platform-level consequences analysed in Chapter 3, the analysis considers three policy documents that have been influential within their respective contexts: the European Union's AI Act (draft dated 19 April 2024) (European Council, 2024), US President Joe Biden's Executive Order on the Safe, Secure, and Trustworthy Development and Use of

Artificial Intelligence (30 October 2023) (The White House, 2023), and Russian President Vladimir Putin's Executive Order 'О развитии искусственного интеллекта в Российской Федерации' [On the development of artificial intelligence in the Russian Federation], together with the 'Национальная стратегия развития искусственного интеллекта на период до 2030 года' [National Strategy for the Development of Artificial Intelligence for the period up to 2030] and the national standards adopted on 15 February 2024 (Kremlin, 2024). Each set of documents reflects its jurisdiction's priorities and institutional constraints, shaping how RS oversight is framed in relation to broader objectives such as economic competitiveness, national security, or information control. The EU instrument selected for comparison is the AI Act because it is the closest functional equivalent to the US Executive Order and the Russian presidential order and strategy as a cross-sector AI governance instrument. This choice necessarily brackets more targeted EU rules that address online platform RS directly, most notably the Digital Services Act, which includes system-level obligations for large platforms and has an emerging implementation record. The analysis therefore treats the AI Act as the EU's general AI governance baseline and uses RS references within that baseline to compare definitional scope, risk framing, and accountability logic across jurisdictions, rather than attempting a comprehensive mapping of all EU instruments relevant to recommender systems.

4.1 Methodology

The analysis examines how each jurisdiction addresses core issues in RS development outlined earlier in this thesis: the cold start problem, algorithmic

bias, economic incentives embedded in RS design, cultural effects of data provenance, and the overall quality of recommendations. This approach highlights how the treatment of these issues reflects underlying value commitments and political priorities, rather than simply exposing uniform deficiencies. In some cases, even in national contexts with strong democratic traditions, regulatory provisions prioritise commercial or political imperatives over explicitly articulated participatory objectives. Particular attention is given to whether RS are explicitly defined or left implicit, and to how selective emphases and embedded ideological assumptions shape their treatment. As in previous chapters, the analysis proceeds across four interrelated dimensions: Technology, Economy, Culture, and Quality (Webster, 2014), enabling consideration of both the stated objectives of the policy documents and the implicit boundaries they establish for RS oversight. The Technology section examines definitions of AI and the ways RS-specific processes are described or categorised. The Economy section addresses compliance costs, innovation pressures, and data governance arrangements. The Culture section interprets underlying normative assumptions and their implications for civic life. Finally, the Quality section focuses on fairness, transparency, and adaptability in RS design and operation. Together, these dimensions enable a structured lexical–discursive analysis of AI policy, positioning RS not only as technical systems but also as epistemic infrastructures relevant to how knowledge is produced, organised, and accessed. To operationalise this comparative framework, all three policy documents were converted into plain text using the Tesseract OCR engine in Python and organised into separate corpora for analysis. The processed text was organised into data frames using the Pandas library,

ensuring consistent formatting across documents and enabling systematic queries and comparisons. This preparation enabled consistent formatting across sources and supported the extraction of comparable features from each text. It also allowed targeted queries on specific terms or sections, facilitating systematic comparison between jurisdictions. The Python package NLTK was applied to preprocessing tasks such as stop word removal, lemmatisation, and generating n-grams (bigrams and trigrams). It was also used for analytical operations including sentiment analysis and measuring association strength. These operations made it possible to identify patterns in evaluative tone and discursive emphasis across jurisdictions. SpaCy was used for tokenisation, extracting prepositional and compound relationships, and identifying keyword co-occurrences. Examining these co-occurrence patterns provides insight into how governance concepts are situated in relation to each other within the discourse and how these relationships differ between regulatory contexts. In addition to lexical and thematic analysis, named entities were extracted using spaCy's `en_core_web_lg` model for English and `ru_core_news_lg` for Russian, with frequency counts recorded for systematic comparison. This process identifies actors, institutions, and other key entities referenced in the texts. Syntactic relationship extraction captures subject–verb–object triples, prepositional links, and compound noun structures, providing insight into how entities and concepts are connected within each jurisdiction's regulatory discourse. Analysing these relational patterns helps to understand not only which governance concepts are present but also how they are positioned in relation to one another. This perspective enables structural comparisons that reveal patterns of emphasis, omission, or linkage that might not be evident from term frequencies alone. Topic

modelling was conducted using Latent Dirichlet Allocation (LDA), which identifies recurring semantic frames that may not be apparent through surface-level frequency counts. The application of LDA was valuable for analysing legal documents, as it can reveal latent themes and support a more context-sensitive interpretation of policy language (Garrido-Muñoz et al., 2021; Rosca et al., 2020). LDA helps address the limitations of relying solely on keyword frequencies, which can give a misleading impression of the prominence of certain terms. Visualisations are generated using Matplotlib to compare the frequency and distribution of selected keywords across corpora. Combining these tools enables a layered reading of the texts, examining both the explicit terminology and the underlying structures through which RS-related concerns are articulated, downplayed, or reframed, allowing connections between thematic emphasis and broader governance priorities to be examined systematically. While these computational techniques provide a valuable means of comparison, applying them to legal and policy texts requires particular care. Unlike social media data or conversational corpora, governance documents are characterised by long sentences, embedded clauses, and specialised terminology that can obscure lexical signals. Translation across languages also introduces variation in word frequency and semantic scope, which can affect pattern recognition. To mitigate these issues, computational findings are interpreted alongside close reading and qualitative coding, ensuring that outputs are not artefacts of tokenisation or OCR processing. This combined approach allows the analysis to retain the advantages of reproducibility and scale associated with NLP while remaining sensitive to the institutional and linguistic features of policy texts.

Sentiment analysis is not used here to infer affective intent or authorial attitude, which would be inappropriate for legal and policy texts, but as a comparative indicator of evaluative framing across corpora. In regulatory discourse, sentiment scores primarily reflect the distribution of modal verbs, normative qualifiers, and evaluative adjectives that signal precaution, aspiration, or permissiveness rather than emotional tone. Applied consistently across jurisdictions, sentiment analysis supports relative comparison of how concepts such as risk, trust, and security are framed, complementing frequency and co-occurrence analysis rather than substituting for close reading. All sentiment outputs are therefore interpreted cautiously and validated against qualitative analysis to avoid over-ascribing meaning to stylistic features of legal drafting.

To examine the cultural dimensions of policy frameworks for AI in the EU, US, and Russia, this chapter uses an integrated framework combining Hofstede's Cultural Dimensions (Hofstede, 1984), Schwartz's Theory of Basic Human Values (Schwartz, 1992), and Fiske's Relational Models Theory (Fiske, 1991). These models provide structured entry points for examining how cultural norms and values are articulated within AI-related legislation in areas such as authority, individual freedoms, community, and risk management, and for analysing how governance texts frame concerns relevant to epistemic agency, without presuming direct links to observed linguistic outcomes. They overlap in their treatment of hierarchical relations, collective versus individual orientations, approaches to uncertainty, and conceptions of fairness or harmony. They were selected because each, in different ways, resonates with cultural features prominent in the jurisdictions examined in this chapter, offering a structured basis

for cross-jurisdictional comparison. Consolidating them into four categories allows for a more coherent and tractable comparison across policy texts, without treating the underlying theories as interchangeable or fully integrated. The category Governance and Authority merges Hofstede's Power Distance, Fiske's Authority Ranking, and Schwartz's Power and Security to examine hierarchical decision-making and legitimacy. Individual vs Collective Orientations integrates Hofstede's Individualism, Fiske's Market Pricing and Communal Sharing, and Schwartz's Self-Transcendence to assess the balance between autonomy and community. Risk and Security combines Hofstede's Uncertainty Avoidance with Schwartz's Security to analyse attitudes towards regulation and risk management. Values and Social Harmony draws on Schwartz's Universalism and Benevolence, Fiske's Equality Matching, and Hofstede's treatment of tradition to understand the cultural underpinnings of societal structure and cohesion. Each category was operationalised through a list of keywords compiled from the terminology used in the original formulations of Hofstede's Cultural Dimensions (Hofstede et al., 2010), Schwartz's Theory of Basic Human Values (Schwartz, 1992), and Fiske's Relational Models Theory (Fiske, 1991). These terms represent the core descriptors employed by the authors when defining their value dimensions and are widely used in comparative studies (Giessner & Van Quaquebeke, 2010; Boer, Berends & Van Baalen, 2011; Haslam, 2004, 1995; Ling & Sheng-Hua, 2016; Cohen, 2009; Wu, 2006; Tung & Verbeke, 2010; Soares, Farhangmehr & Shoham, 2007; Manrai & Manrai, 2011). Synonyms and morphological variants in both English and Russian were added to improve coverage of the policy texts. The full keyword lists for each category are presented in Table 1. This consolidation ensures that the cultural analysis

captures both the distinct emphases of each framework and the areas where their concepts converge.

Table 1 Keyword sets for the combined cultural framework categories.

English	Russian
Governance & Authority	
authority, hierarchy, control, power, leadership, compliance, obedience, tradition, preservation, modesty, conformity	власть, иерархия, подчинение, контроль, руководство, традиции, скромность, соблюдение, сохранение
Individual vs Collective Orientations	
individual, personal, privacy, self, autonomy, person, community, group, responsibility, obligation, sharing	индивидуальность, личный, конфиденциальность, сам, автономия, общество, группа, ответственность, распределение
Risk & Security	
security, safety, stability, protection, future, persistence, prudence, environment, daring, ambition, exploration	безопасность, защищенность, стабильность, осторожность, будущее, настойчивость, окружение, дерзость, амбиции, исследование
Values & Social Harmony	
fairness, justice, equality, balance, reciprocity, peace, harmony, benevolence, sharing	справедливость, правосудие, равенство, равноправие, баланс, взаимность, мир, гармония, доброжелательность, обмен, пространство

Note: "Risk" is not included as a keyword because it functions as an umbrella term across categories in policy discourse and would inflate counts without improving interpretive specificity; the category instead uses security- and safety-related terms as proxies for how risk is substantively framed.

Combining multiple theories can reduce granularity, so original dimensions are revisited when necessary. Cultural specificity in legal terminology can create challenges for cross-context interpretation, which is addressed through systematic synonym expansion. The expanded lists were created by taking the original keywords from Hofstede, Schwartz, and Fiske and extending them with semantically related terms drawn from WordNet for English and RuWordNet for Russian. For instance, in the Governance and Authority category, the core term ‘power’ was supplemented with synonyms such as ‘dominance,’

‘control,’ and ‘authority,’ while in Values and Social Harmony, the descriptor ‘fairness’ was broadened with related terms such as ‘justice,’ ‘equality,’ and ‘reciprocity.’ This ensured that differences in culturally specific terminology across policy texts were captured within the same conceptual categories, reducing the risk of missing semantically equivalent but lexically distinct expressions. The interpretive nature of merging frameworks may introduce researcher bias, which is mitigated through triangulation with qualitative content analysis and comparison across multiple computational techniques, rather than eliminated. Throughout this chapter, regulatory texts are treated as discursive artefacts that articulate priorities, assumptions, and governance rationales, rather than as direct indicators of implemented policy or system behaviour. While these documents influence legal and institutional practices, they do not guarantee execution or uniform enforcement. The analysis does not assume that provisions in the EU AI Act, Biden’s Executive Order, or Putin’s AI Strategy translate directly into technical design choices or platform behaviour, focusing instead on how these texts construct meaning, define priorities, and encode assumptions relevant to RS governance. The cultural frameworks are used as interpretive aids rather than fixed taxonomies, supporting inference about how regulation was framed without making claims about national character or public opinion. Nature Language Processing tools such as topic modelling, co-occurrence, and sentiment analysis enable systematic pattern recognition but require researcher validation to avoid misinterpretation of legislative intent. Keyword frequency and association strength are interpreted within the specific institutional and political contexts of each document. All computational findings are cross validated with qualitative content analysis, with results from code

checked through manual coding techniques and LancsBox software to confirm their accuracy. With these methodological foundations established, the next step is to examine how the regulatory texts conceptualise technology, beginning with the definitions of AI that frame RS within each jurisdiction's policy discourse and linking these framings to the technical risks identified in Chapter 2.

4.2 Technology: To what extent do AI definitions in the EU, US, and Russia explicitly frame RS as a distinct technical object?

The explicit and implicit definitions of AI within the legislative documents analysed in this chapter play an important role in shaping the scope and application of the proposed frameworks, influencing their relevance to RS development (Thierer, Castillo O'Sullivan & Russell, 2017). What counts as 'technology' in policy or regulatory discourse reflects underlying assumptions about what should be governed, the optimal mechanisms for doing so, and the actors responsible. In AI regulation, 'technology' is often framed at a high level of abstraction, encompassing software techniques, cognitive emulation, and predictive outputs, yet it is rarely disaggregated into distinct system types. RS are therefore often subsumed into broad AI categories, limiting opportunities for targeted oversight despite their significant role in structuring information environments. Schuett notes that policymakers 'should not use the term "artificial intelligence" for regulatory purposes' because definitions vary substantially across disciplines and lack precision (Schuett, 2019). The problem is particularly relevant for RS, whose foundational techniques, including collaborative filtering, content-based filtering, hybrid systems, deep learning, loss functions and context-aware approaches, span multiple domains of AI development. As

discussed in Chapter 2, these systems present technical and ethical risks that are not adequately addressed when regulatory definitions treat them as generic AI. This also reflects a wider ‘pacing’ problem in technology governance, where regulation can lag behind technical change and detailed technical prescriptions risk rapid obsolescence. For systems such as RS, specifying mechanisms at the level of LF, cold start strategies, or particular model classes can invite compliance-by-framing, where regulated actors define the operative meaning of technical terms and shift implementations while claiming formal conformity. For this reason, regulatory practice often favours technology-neutral definitions and obligations stated in functional terms, complemented by standards, guidance, and auditing practices that can be updated more quickly than primary legislation.

The cold start example illustrates why detailed technical definitions can weaken regulatory robustness. Features such as cold start are not stable system properties but socio-technical constructs whose relevance, mitigation, and impact are largely shaped by platform practices and commercial incentives. When regulation anchors obligations to specific technical mechanisms, regulated actors retain significant discretion to redefine compliance through architectural or procedural shifts, creating conditions for regulatory capture. This dynamic is well documented in research on technology-neutral regulation (Hildebrandt & Tielemans, 2013; Puhakainen & Väyrynen, 2021; Koops, 2006; Marchant, 2011), which argues that functional and outcome-oriented obligations are more resilient to rapid technical change than implementation-level prescriptions. The evolution of the EU AI Act itself reflects this dilemma: the later introduction of general-purpose AI provisions in response to large language models illustrates how

legislative scope must adapt to emergent system classes without locking governance to transient technical features (Veale & Zuiderveen Borgesius, 2021). Read in this context, the limited technical specificity applied to RS reflects not only omission but a structural trade-off between precision and durability in AI governance.

Without explicit framing in policy documents, RS-specific concerns are often left without targeted regulatory attention. These include the cold start problem, where limited user data at entry can lead developers to adopt intrusive data collection practices, algorithmic bias that may arise from optimisation strategies prioritising engagement metrics, and opacity that makes RS outputs difficult to interpret or contest. These risks have downstream effects on user experience and civic participation. For instance, when cold start strategies rely on demographic proxies or third-party tracking, the parameters of epistemic agency are influenced before any meaningful interaction occurs. Bias embedded in loss functions can reproduce existing social inequalities, and opaque RS designs reduce the ability of users to understand or question recommendation logic. Because RS can shape information access and the conditions of public discourse, their absence from AI regulatory definitions limits the scope for addressing these impacts in a systematic way. This raises questions about whether abstract AI categories are sufficient for capturing the design and operational characteristics of RS within existing regulatory frameworks.

The definitions of AI in the EU AI Act, Biden's Executive Order, and Putin's EO differ in emphasis, reflecting the regulatory priorities of their respective

jurisdictions. The EU AI Act focuses on the techniques underpinning and outputs of AI systems, with particular attention to accountability for the processes that shape their development. It proposes ‘a single future-proof definition of AI,’ while acknowledging that ‘stakeholders mostly requested a narrow, clear and precise definition for AI’ (European Parliament, 2024). The analysis here uses the publicly available draft text dated 19 April 2024. The final Act in force retains an OECD-aligned, outcome-oriented framing of AI that is closer to the US statutory definition, and this does not change the comparative point that RS remain treated as a generic capability rather than a distinct system type. The legislation defines an ‘artificial intelligence system’ as:

... software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with. (European Parliament, 2024: p. 39)

This definition is supported by Annex I which consists of a list of techniques:

**ANNEX I: ARTIFICIAL INTELLIGENCE TECHNIQUES AND
APPROACHES**
referred to in Article 3, point 1

- (a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;
- (b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
- (c) Statistical approaches, Bayesian estimation, search and optimization methods. (EUR-Lex, 2021)

This outcome-oriented definition parallels the EU framework in scope but differs by explicitly acknowledging human contributions in addition to machine outputs, thereby more accurately reflecting the complexity of assigning accountability across both technical design choices and user-generated inputs. The EU AI Act adopts a risk-based framework categorising AI applications into four tiers: minimal or no risk, limited risk, high risk, and unacceptable risk. RS are placed in the minimal or no-risk category, based on the view that they generally process non-sensitive information obtained through user consent mechanisms such as cookies or browser settings, though indirect profiling or inference remains possible. By contrast, technologies such as facial recognition or behaviour manipulation, classed as unacceptable risk, involve data types and processing modes with a higher potential to harm legal interests (Schuett, 2023). High-risk systems, including those in transportation, safety, and access to education or employment, may process similar data to RS but are subject to stricter oversight due to their recognised societal impact. This low-risk classification of RS underscores the difficulty of aligning legal definitions with the broader social and cultural concerns discussed in the literature, as these systems are associated with patterns of information exposure that extend beyond narrowly defined risk categories. Without provisions for such effects, systems deemed low risk may operate outside direct regulatory scrutiny, allowing subtle forms of cultural and political influence to develop gradually and remain difficult to reverse (Helberger, Pierson & Poell, 2018).

The Executive Order issued by the Biden administration frames AI differently, focusing on the impact of a technological system rather than on technical processes or outputs alone:

The term “artificial intelligence” or “AI” has the meaning set forth in 15 U.S.C. 9401(3): a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action. (The White House, 2023)

Biden’s EO builds on the National Artificial Intelligence Initiative Act and prioritises the societal impact of AI in specific domains over the type of data it processes (Rep. Johnson, 2020). In contrast to the EU AI Act’s tiered, data-centric classification, the US approach evaluates AI according to its influence on critical infrastructure and key societal functions. This framing delegates responsibility to federal agencies to identify, assess, and mitigate risks within their respective mandates. Examples include directives to evaluate vulnerabilities in energy grids or communications networks and to develop targeted mitigation strategies. In this structure, RS may be subject to oversight from multiple agencies operating under different guidelines. This can lead to overlapping jurisdictions or, conversely, regulatory gaps when no agency claims responsibility for a technology in a given domain. Agencies regulate AI according to its operational environment, as in the Food and Drug Administration’s remit over medical applications. While this decentralisation supports adaptability and responsiveness to emerging uses, it can also delay oversight and fragment access to expert knowledge, which limits the capacity to integrate informed

perspectives and reduces the system's ability to anticipate and futureproof against emerging risks. For instance, convolutional neural networks (CNNs) originally designed for image classification in photography later found application in X-ray analysis, with FDA approval processes triggered only after such medical uses emerged. A further example is CNN application in WaveNet, initially developed for generating naturalistic speech and trained on acoustic datasets, which was subsequently integrated into Google's Duplex system to enable conversational voice interactions (Leviathan & Matias, 2018). Cross-domain adaptations such as this illustrate how technical innovation often precedes regulatory consideration. They also highlight the challenge of a reactive framework in anticipating the societal implications of AI systems, particularly when they enter high-stakes contexts before oversight mechanisms are in place.

The definition of AI in President Putin's EO and the accompanying National Strategy emphasises cognitive mimicry rather than technical methods or measurable outputs:

комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые с результатами интеллектуальной деятельности человека или превосходящие их. Комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру, программное обеспечение (в том числе в котором используются методы машинного обучения), процессы и сервисы по обработке данных и поиску решений. [A set of technological solutions that allows imitating human cognitive functions (including problem-solving without a pre-defined algorithm) and obtaining results when performing specific tasks comparable to or exceeding the results of human intellectual activity. The set of technological solutions includes information and communication infrastructure, software (including those using

machine learning methods), data processing and problem-solving processes, and services. Author's translation] (Kremlin, 2024)

Russia's definition frames AI in terms of its ability to replicate human-like cognitive capabilities, prioritising functions and decisions that resemble human reasoning over considerations such as data sensitivity or domain-specific application. This orientation aligns with strategic objectives of safeguarding national security and reinforcing state ideology (Barnes, 2024; Sonne & Das, 2024). AI is positioned as an instrument of state sovereignty, with development and deployment subject to direct governmental oversight across all sectors (Nitsevich et al., 2020; Jennings, 2021). Within this framing, RS are understood not merely as content delivery tools but as systems capable of simulating decision-making processes and analysing user preferences, enabling their integration into state-controlled information environments with functions that extend beyond personalisation. This perspective focuses on the intended role of the technology rather than its underlying mechanics or output metrics. While such an approach allows for flexible integration into different domains, it leaves open critical questions about bias mitigation, privacy safeguards, and transparency. The analogy to human actors illustrates this gap: when a salesperson recommends clothing, their preferences and assumptions are typically accepted without formal scrutiny, and there is no expectation of disclosing every influencing factor. In the digital sphere, however, the scale, persistence, and systemic embedding of such biases can produce cumulative effects that differ substantially from their human counterparts, warranting more deliberate oversight.

Across all three jurisdictions, the absence of explicit recognition for RS limits the capacity of policy frameworks to address the technical and societal challenges these systems present. The contrasting approaches to defining AI, technique-based in the EU, output-based in the US, and centred on cognitive mimicry in Russia, shape both the scope of downstream regulation and the mechanisms through which RS are ultimately governed. These definitional choices influence how RS are designed, deployed, and embedded in public life. As discussed in Chapter 3, platforms such as YouTube and Dzen provide empirical contexts in which RS design decisions and operational priorities can be examined alongside the absence of RS-specific regulatory provisions. Understanding how these definitions intersect with the technical characteristics of RS is essential for examining the specific challenges explored in this dissertation.

Definitions are not neutral descriptors but legal instruments that establish jurisdiction and allocate responsibility. In the EU context, for example, the interpretation of 'personal data' under the General Data Protection Regulation has determined the extent of compliance obligations across sectors (Schuman, 2024). A similar dynamic applies to RS: when they are not explicitly defined, regulators may lack the mandate to investigate their effects on information environments, even if their societal impact is substantial. Definitions also influence which systems trigger oversight, which agencies can intervene, and what kinds of accountability measures are expected. Treating RS as incidental functions of broader AI categories risks leaving key technical practices, such as ranking or optimisation, outside the scope of enforceable standards. The way AI

is defined therefore shapes not only the scope of regulation but also the epistemic conditions under which RS operate in civic and political life.

RS-specific technical issues

RS raise technical issues, including cold start, optimisation-driven bias, echo chambers, and overfitting, yet governance texts typically approach these systems through technology-neutral duties rather than prescribing implementation-level techniques. This reflects a trade-off: detailed prescriptions can increase immediate specificity, but they can also become outdated and create opportunities for strategic compliance as RS architectures and business practices evolve. Against this backdrop, the discussion below reads the three frameworks as setting general constraints that may indirectly shape RS practice, while noting where RS-relevant risks are not clearly captured by those constraints. For instance, the EU AI Act requires disclosure of data collection practices and offers users control over their information, yet it does not outline strategies for mitigating cold start effects in systems that influence civic participation. In the absence of such guidance, developers may continue using data-gathering approaches that shape user profiles before meaningful interaction occurs, influencing early-stage exposure to political or civic content. Algorithmic biases, particularly those linked to the optimisation criteria embedded in loss functions, also shape RS outputs in ways that can reinforce systemic inequalities. As discussed in Chapter 2, these biases often prioritise engagement over fairness, amplifying content that maximises clicks or watch time. None of the frameworks analysed makes explicit reference to loss functions or similar implementation-level choices, consistent with a preference for technology-

neutral drafting. The EU AI Act prohibits discriminatory outputs and includes fairness requirements, but its enforcement mechanisms are directed mainly at high-risk systems, leaving RS outside their main scope. Biden's EO refers to fairness in general terms but does not specify measures aimed at optimisation strategies. In Russia's framework, the emphasis on technological sovereignty and state control over AI development does not extend to addressing algorithmic accountability or bias mitigation. Echo chambers and filter bubbles, as detailed in Chapters 2 and 3, represent another unaddressed RS-specific issue. These effects occur when RS continually reinforce existing preferences, limiting exposure to diverse viewpoints and narrowing the informational environment. In political contexts, this can affect epistemic agency by reducing the breadth of content available to users. The EU AI Act recognises risks related to manipulation but provides few safeguards against echo chamber formation. Although it prohibits systems that 'significantly distort democratic processes,' it does not define operational measures for maintaining diversity in RS recommendations (European Council, 2024). Biden's EO mentions misinformation but leaves its mitigation to voluntary industry standards. In the Russian context, RS are framed as tools for advancing state objectives, a perspective that can structurally embed ideological alignment in recommendation design. Chapter 3's case study of Dzen illustrates how such alignment can occur, with content clustered around narratives consistent with official positions. Overfitting, another technical challenge described in Chapter 2, reduces RS adaptability by tailoring recommendations too narrowly to historical data. This can result in repetitive content streams and limit the introduction of new or unexpected material. Despite its implications for informational diversity, none of the policy frameworks includes

provisions that directly address overfitting or strategies for balancing personalisation with generalisation. The EU AI Act requires some high-risk systems to disclose their decision-making logic, yet RS fall into the minimal- or low-risk categories and are excluded from these requirements. Biden's EO promotes the development of 'trustworthy' AI but does not include specific evaluation criteria for overfitting in RS. In Russia's framework, efficiency and alignment with state-defined objectives take precedence, leaving the risks of over-optimisation unexamined.

These omissions have identifiable downstream effects. Opacity in RS, particularly those built on deep learning or hybrid architectures, constrains public awareness of the mechanisms that shape recommendations. As shown in Chapter 3, YouTube's engagement-focused model filters content through parameters that remain undisclosed to users, producing information streams that can narrow perspective without overt indication of curation. Optimisation choices, such as those embedded in loss functions, structure outputs in ways that reproduce demographic targeting patterns, influencing what types of content different groups encounter. Chapter 3 illustrated how, in both YouTube and Dzen, minimal early interaction data was sufficient to trigger delivery of highly polarised political material to users. These dynamics affect epistemic agency by limiting opportunities to encounter, assess, and integrate diverse perspectives. When selection criteria are inaccessible to the user, the ability to interrogate or reframe one's information environment is reduced. In the absence of provisions tailored to RS, design strategies continue to reflect platform-level priorities such as engagement or retention, with the technical issues outlined in Chapter 2, such

as cold start, bias, echo chambers, and overfitting, remaining outside specific regulatory focus. Across the three governance frameworks, RS are incorporated only under broad AI categories, leaving unaddressed the ways in which their architecture and operational logic influence access to political information. As a result, the forms of influence documented in Chapter 3 persist without direct regulatory intervention, reinforcing the informational conditions that the policy frameworks imply to manage.

Examples

To examine how issues are positioned within governance discourse, a frequency comparison was conducted for the terms ‘recommendation,’ ‘bias,’ and ‘transparency’ across the three policy corpora analysed in this chapter, with attention to the fact that each term can denote distinct and sometimes divergent concepts depending on regulatory or academic context. These terms were chosen for their direct relevance to RS design and their prominence in both regulatory and academic debates on AI governance. Figure 10 visualises their normalised per 10,000 words frequency in each corpus, revealing jurisdictional variation in how these concepts are framed.

In the corpus, ‘recommendation’ appears with the highest relative frequency in Biden’s EO, where it is consistently framed in its normative sense as a mechanism for guiding institutional practice across domains. In this analysis, the term ‘recommendation’ is used in two distinct senses: normatively, to denote advisory or coordinative guidance within governance processes, and technologically, to denote algorithmic outputs generated by RS. In the EU AI Act,

the term is less prominent and is primarily associated with safeguards and human oversight mechanisms. References to the reversibility of AI outputs, such as ‘predictions, recommendations or decisions of the AI system can be effectively reversed and disregarded,’ frame recommendations as outputs that must remain subject to human control. In Putin’s EO, ‘recommendation’ occurs proportionally less often than in the US corpus but more frequently than in the EU, where it is used in administrative contexts to emphasise coordination and the dissemination of best practices.

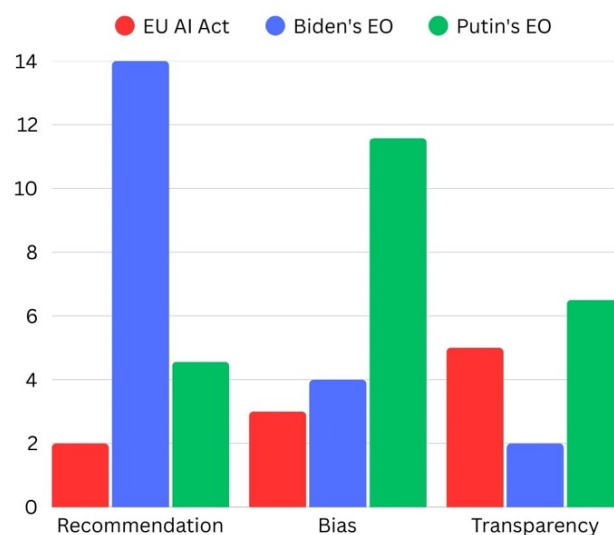


Figure 12 Normalised per 10k words frequency comparison for the recommendation, bias, and transparency.

Similarly, the provision to ‘override a decision or recommendations that may lead to potential harm’ connects to the governance principle that high-stakes outputs must be contestable. While RS are classified as minimal or no-risk, the Act’s discussion of ‘high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons’ reflects an awareness of automation bias, where over-reliance on system output may reduce critical engagement. This has implications for how RS are discussed in policy, particularly in relation to concerns about over-reliance on automated

outputs in decision-support contexts. In the US policy, the term serves primarily in definitional and policy advisory contexts. The clause ‘make predictions, recommendations, or decisions influencing real or virtual environments’ forms part of the overarching AI system definition, situating recommendations as a generic function rather than a distinct technical process. Its appearance in ‘issue recommendations to the President on potential executive actions relating to copyright and AI’ reflects a procedural, non-technical sense of the term, relevant to policymaking rather than RS design. In the Russian corpus, ‘recommendation’ appears in technical and regulatory contexts that directly name RS. The reference to ‘правил регулирования рекомендательных алгоритмов в Интернете’ [rules for regulating recommendation algorithms on the Internet] acknowledges formal rulemaking for internet-based recommendation algorithms. The phrase ‘в рекомендательных системах, тестируемых без участия пользователя’ [in RS tested without user participation] introduces considerations for model evaluation without user interaction. The example of ‘рекомендательная система, которая отслеживает реакцию пользователя на свои рекомендации’ [RS that tracks user response to its recommendations] points to adaptive systems that adjust outputs based on user behaviour, raising questions about autonomy and the potential for unmonitored influence. Across all three policies, the treatment of ‘recommendation’ illustrates a conceptual gap. Even where technical or regulatory awareness exists, RS-specific dynamics across domains and applications remain underdeveloped, limiting the ability to address their cumulative societal effects.

The EU text's requirements for human oversight, reversibility, and contestability point toward technical affordances that RS could implement but rarely operationalise in practice. Logging recommendation paths would allow for post-hoc examination of how a specific output was generated, model versioning would enable rollback to earlier states in the event of harmful drift, and counterfactual generation could help users and auditors see what alternatives the system might have produced. The human override clauses align with interrupt mechanisms that could pause or reset recommendation flows when they appear to be steering users toward harmful or misleading content. As discussed in Chapter 3, automation bias is a relevant concern here, as users often accept early RS outputs without interrogation. In YouTube's case, observations from Chapter 3 illustrate how early recommendation outputs can become clustered, providing an empirical context for why automation bias is treated as a policy concern. While such effects emerge gradually, their accumulation can restructure the informational environment in ways that are difficult to reverse without architectural changes to the RS itself. The limited operationalisation of these affordances in governance texts, and the absence of procedural detail in the US and Russian materials, underscores a broader pattern: even when recommendations are mentioned explicitly, they are rarely connected to the technical levers that could manage their influence. This framing reinforces the technological point made previously that RS are subsumed under broad AI categories, leaving the specific design choices and operational behaviours of these systems outside targeted governance.

In the corpus, 'bias' generally appears more frequently than 'recommendation' but is framed in ways that limit its direct application to RS. In the EU AI Act, the term is linked to institutional impartiality, as in 'exercise their powers independently, impartially and without bias,' where it refers to the conduct of authorities rather than the properties of systems. It is also present in data governance provisions that require 'methods to detect identifiable biases, where applicable' during curation. While these measures could be relevant to RS, they are not designed to address the specific ways in which recommendation algorithms inherit or amplify bias through training data selection, similarity scoring, or optimisation strategies. The reference to 'technical inaccuracies... can lead to biased results and entail discriminatory effects' addresses biometric identification systems classified as high risk, with no parallel specifics given for RS, even though similar optimisation patterns can produce skewed outputs in recommendation contexts. In the US policy, bias is treated primarily as a compliance concern within regulated sectors. The directive to 'evaluate their underwriting models for bias or disparities affecting protected groups' places the focus on financial decision-making. Instructions to 'minimize bias' in 'automated collateral-valuation and appraisal processes' address real estate valuation, and calls for 'monitoring algorithmic performance against discrimination and bias in existing models' apply to healthcare systems. These measures show operational bias mitigation, yet none engage with RS-specific interventions such as ranking adjustments, diversity quotas in output lists, or post-processing to counter selection bias. The Russian corpus combines conceptual and operational references. The statement 'справедливость – это не то же самое, что отсутствие предвзятости' [fairness is not the same as the absence of bias]

draws a distinction between fairness and bias, positioning bias as a broader ethical and cognitive phenomenon. The glossary entry for ‘когнитивная предвзятость’ [cognitive bias] focuses on human judgment, while mentions of ‘несправедливой предвзятостью’ [unfair bias] and ‘потенциальной предвзятости, ошибок, или вреда’ [potential bias, errors, or harm] incorporate bias into risk assessments for AI deployment. Although these definitions recognise bias at both human and system levels, they do not address the specific mechanisms by which RS propagate it. Across all jurisdictions, bias is acknowledged as a governance concern but remains addressed through general principles or domain-specific rules, with RS-specific sources of bias and potential mitigation strategies largely absent from the regulatory discourse. The treatment of bias across jurisdictions supports the observation that general regulatory principles do not explicitly engage with the technical pathways through which bias can emerge in RS architectures.

In the EU AI Act, ‘transparency’ is embedded in sector-specific obligations. The reference to how ‘accuracy, reliability and transparency is particularly important to avoid adverse impacts’ addresses AI systems in law enforcement, where performance standards and public trust are directly linked. In the regulation on ‘transparency and targeting of political advertising,’ the focus is on ensuring that audiences can identify the origin and intent of political messages. The requirement for ‘transparency about the original purpose of the data collection’ in data governance provisions reflects an emphasis on compliance with privacy law. While these measures touch on information flows and public accountability, they do not extend to making the decision-making

processes of RS visible or understandable to users. This omission is notable given the role RS play in structuring exposure to political content and other forms of targeted messaging, as discussed in existing research. In the US policy, transparency is primarily an institutional obligation. The directive to ‘improve transparency for agencies’ use of AI’ calls for reporting on government AI deployments, aiming to make system uses and associated risks more visible to the public. In the workplace context, ‘transparency, engagement, management’ in relation to ‘employers’ AI-related collection and use of data’ establishes expectations for informing employees about monitoring and data handling. References to the ‘transparency of AI models’ form part of broader regulatory expectations, but no procedures are outlined for RS-specific disclosures, such as explaining ranking criteria or personalisation parameters.

The Russian corpus contains both conceptual and procedural framings. The definition ‘прозрачность: свойство системы искусственного интеллекта’ [transparency: property of an artificial intelligence system] describes transparency as a property of AI systems, while ‘прозрачности и объяснимости систем ИИ’ [transparency and explainability of AI systems] positions it as an aspect of reliability. The principle of ‘обеспечение прозрачности целей и процессов организации’ [ensuring transparency of organisational goals and processes] extends transparency to organisational objectives and workflows. These definitions emphasise clarity and accessibility of information but do not prescribe technical methods for revealing the internal logic of RS or documenting their adaptive behaviours. Across the three jurisdictions, transparency is recognised as essential for legitimacy and trust in AI, yet its application to RS

remains underdeveloped. Sector-specific or conceptual framings dominate, leaving limited guidance on how transparency might be operationalised for recommendation processes that structure patterns of information access. The absence of RS-specific transparency provisions reflects the broader issue raised previously that technical affordances central to RS operation remain unaddressed when governance is framed in abstract AI terms.

Taken together, the frequency data and the extracted excerpts indicate that RS occupy a peripheral position in all three governance frameworks. Their technical distinctiveness and the concerns identified in Chapters 2 and 3 are not reflected in how these terms are operationalised in legislative or strategic texts. This discursive marginality is not simply a matter of omission but reflects the way in which RS are linguistically and conceptually subsumed under the broad category of AI. The next section considers the economic dimensions of this marginality, focusing on how regulatory approaches intersect with commercial pressures, compliance costs, and data governance in shaping RS design and operation.

4.3 Economy: How do the economic priorities implicit in AI policies influence the regulation of RS?

The economic implications of AI regulation shape multiple stages of RS development, influencing how compliance requirements are balanced with innovation goals and market competitiveness. Frameworks such as the EU AI Act and Biden's Executive Order emphasise the functioning of the internal market and the promotion of a fair, open, and competitive AI ecosystem. These

objectives can increase development costs through obligations such as detailed documentation, additional data pre-processing, and adjustments to system outputs. Such measures influence how developers allocate resources across training, testing, and deployment. As discussed in Chapter 2, RS are frequently designed to maximise engagement in contexts where commercial imperatives determine technical priorities and long-term strategies. This economic logic incentivises optimisation for metrics such as click-through rates and targeted content delivery, often reinforcing existing user preferences at the expense of fairness and diversity. The EU and US frameworks both seek to balance regulation with competitiveness, though they approach this balance differently. In the EU, the aim is to create a transparent and predictable market for investors, reducing uncertainty while imposing significant compliance obligations. Biden's Executive Order emphasises sustaining innovation momentum by lowering regulatory barriers and encouraging collaboration between government and industry. Putin's Executive Order departs from market-oriented models entirely, centralising AI development under state authority to align innovation with national objectives and limiting the private sector's role. Across these approaches, regulation shapes the conditions under which RS are developed and deployed. This influence can be seen in how investment environments are structured, how data is positioned as a monetizable asset, and how innovation priorities are directed. Each framework embeds a specific economic logic that shapes regulatory incentives and constraints relevant to how RS are developed and deployed.

Economic compliance burdens vary across jurisdictions and shape the conditions in which RS are developed. The EU AI Act requires extensive documentation, monitoring, and data retention, which impose significant costs, particularly on small and medium-sized enterprises. These demands disproportionately affect startups with limited resources, reducing market diversity and consolidating innovation among larger, established developers (Cath et al., 2017; Cennamo et al., 2023; Geradin, Bania and Karanikioti, 2022). The EU framework builds on a regulatory tradition shaped by the General Data Protection Regulation (GDPR), which explicitly links personal data to monetary value and recognises it as a substitute for payment in the digital economy (Malgieri and Custers, 2018). This framing positions data as a core market asset underpinning the supply chain for digital content and services. The AI Act strengthens this link through dataset documentation requirements aimed at transparency, security, and accountability. These measures address risks such as cyberattacks, data poisoning, and dataset quality issues, yet their associated costs create additional barriers for new entrants and contribute to regulatory concentration among larger firms (Ebers et al., 2021; Wischmeyer and Rademacher, 2020; Lilkov, 2021). Biden's Executive Order adopts a more flexible model, encouraging public-private partnerships and sector-specific collaboration. It promotes AI development through data sharing, education, intellectual property protection, and rapid market integration. This environment lowers entry barriers and accelerates deployment, but it leaves ethical risks largely unmanaged in early development, placing responsibility for fairness and transparency on private actors. This approach facilitates experimentation while leaving issues such as optimisation transparency, data governance, and

alignment with civic objectives less explicitly specified at the regulatory level. As discussed in Chapter 2, it creates conditions in which RS are optimised for engagement at the expense of fairness, often amplifying demographic targeting and filter bubbles. Russia's AI Strategy follows a different logic. All AI-related data must be stored on government-controlled infrastructure, eliminating private-sector data governance. This centralisation shifts data management costs to the state but removes developer autonomy, requiring alignment with government-defined priorities. The model limits private-sector involvement, restricts international collaboration, and isolates the AI ecosystem from global data flows (Epifanova, 2020; Kostyuk and Zhukov, 2019). While financial burdens on developers are low, political control is entrenched and RS adaptability to diverse or critical inputs is constrained.

Cross-border data flows highlight the divergence between the three models. In the EU, strict data protection rules complicate international transfers, requiring developers operating across jurisdictions to comply with overlapping storage, localisation, and audit obligations that increase costs. The US enables more open data exchange through commercial platforms and research partnerships, but without equivalent privacy safeguards. Russia prohibits cross-border transfers altogether, reinforcing national sovereignty while reducing interoperability and limiting access to global datasets. These differences extend to the investment environment. The EU AI Act offers a transparent and predictable regulatory framework that can attract investors through clear compliance expectations, yet its cost-intensive requirements limit the viability of high-risk or experimental projects (Massey and Machin, 2023). In the US, lighter

regulatory burdens and minimal front-loaded compliance allow faster development cycles and wider market participation, though the absence of mandatory accountability increases longer-term legal and reputational risks. Russia's framework removes market competition and directs AI development through state institutions, constraining private-sector innovation and embedding inefficiencies into national infrastructure (Zhdaneev, 2022). Data governance is especially consequential for RS development because the rules governing data storage, transfer, and use determine which actors can build such systems and what resources they can access. The economic design of regulatory frameworks conditions the environments in which RS are developed, with compliance costs, entry barriers, and state controls influencing the range of systems that reach deployment. In high-cost compliance environments, development is concentrated among well-capitalised actors who can absorb documentation and audit obligations. This reduces market diversity but may support more consistent oversight. In lighter regimes, lower entry barriers encourage experimentation and a wider variety of system architectures, but monitoring is weaker and optimisation for engagement can proceed unchecked. In state-controlled models, market competition is removed, and RS outputs are aligned with political priorities, limiting exposure to dissenting or minority perspectives. Each model thus embeds a trade-off between innovation, accountability, and plurality. These trade-offs shape the regulatory and market conditions under which recommendation systems operate, with implications for the diversity and repetition observed in recommendation outputs. As discussed in Chapter 2, RS depend on user data to produce personalised outputs. Whether collected through cookies, platform interactions, or third-party tracking, its storage, use, and

processing set the boundaries of ethical RS design. In the EU, collection is legally constrained and linked to user consent. In the US, emphasis on data availability and partnership encourages broad aggregation and reuse, sometimes without user awareness. In Russia, all AI-related data is processed through government-controlled channels, enabling the embedding of ideological priorities into RS training and output, as illustrated in Chapter 3's analysis of Dzen. These economic logics shape not only who can build RS but also what these systems optimise for: in high-cost compliance environments, well-funded actors dominate; in flexible regimes, engagement takes precedence over civic alignment; in state-controlled models, optimisation follows political priorities rather than public interest.

Analysis of the three corpora, shown in Figure 11, uses raw frequencies to reflect how often these economic terms appear in each text, confirming the distinctions identified across jurisdictions. In the EU AI Act, 'market' appears 423 times compared to 156 occurrences of 'compliance,' indicating a consistent focus on market functioning alongside regulatory requirements. In Biden's Executive Order, occurrences of these economic terms are minimal: 'market' appears 11 times and 'compliance' 8, while 'cost' and 'investment' occur 16 and 5 times respectively, reflecting its emphasis on resource allocation and growth incentives. In Putin's Executive Order, compliance is expressed mainly through the term 'соблюдение' [compliance] which dominates overwhelmingly, while 'рынок' [market] is referenced only 6 times, signalling a regulatory vocabulary tied to state control rather than competitive markets. In the secondary set of terms, Russia also leads in mentions in raw frequencies of both 'cost' and

‘investment,’ though again framed in state-led development contexts rather than open-market activity. These distributions support the structural finding that while the EU balances regulatory oversight with market participation and the US foregrounds resource deployment over formal compliance, Russia frames AI primarily as a strategic infrastructure project. The frequency patterns indicate how economic priorities are articulated within policy texts, providing contextual signals relevant to how RS development is framed within each jurisdiction. In Russia, data localisation requirements lead to domestically hosted infrastructures, while in the EU, compliance documentation encourages modular designs that support audit trails. These patterns show that economic drivers of RS design extend beyond regulation itself, reflecting platform incentives, market strategy, and the wider commercial environment in which these systems operate. As discussed in Chapter 2, RS are built to maximise outputs such as click-through rate, time on platform, and repeat interaction, and these metrics directly influence both architectural choices and the optimisation priorities embedded in their operation.

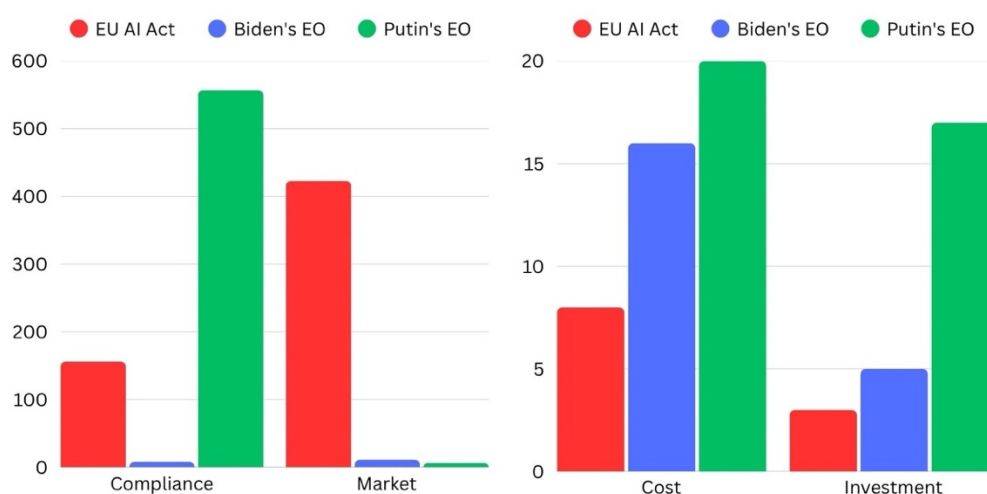


Figure 13 Raw frequency of economic terms compliance, market, cost, and investment.

A further economic factor influencing RS design is the strategic use of architectural opacity. Black-box systems, including deep learning and multi-layer neural networks, dominate large-scale RS because they support complex pattern recognition and personalised ranking at scale. In these systems, techniques such as collaborative filtering and hybrid modelling integrate numerous input streams, creating outputs that are technically reproducible but practically resistant to external interpretation. This resistance is not solely a by-product of complexity. It can be an intentional design choice, maintaining proprietary control over recommendation logic and insulating the system from outside scrutiny. For monetised platforms, this opacity creates advantages beyond the technical domain. It sustains advertiser confidence in the precision of targeted delivery while limiting the ability of third parties to detect ranking biases or evaluate the breadth of content exposure. On YouTube, for example, feedback loops develop when early interactions are amplified through iterative ranking, concentrating the user's informational field. In Dzen, a similar narrowing occurs, though driven by alignment with state-approved content streams. In both cases, the closed nature of the architecture obscures the mechanisms behind content selection, making it difficult for users to identify or challenge embedded priorities. These design choices also reduce the likelihood that platforms will introduce features enabling recommendation audits, such as model version histories or counterfactual explanations. From an economic standpoint, such features can be seen as introducing risk to engagement metrics by making targeting logic more transparent and, therefore, contestable. This absence of technical accountability mechanisms persists across the three governance models examined. In each case, the economic logic underpinning RS development is compatible with

opaque optimisation practices, which can make recommendation pathways difficult to scrutinise and adjust.

The compliance structures described within each policy structure add another layer to these economic dynamics. In the EU AI Act, compliance is framed not only as a safeguard but as a mechanism for market integration. Standardisation is presented as a means to ‘promote innovation as well as competitiveness and growth in the single market,’ with harmonised standards serving as evidence of conformity. This is reinforced by the regulatory sandbox framework, designed to ‘enhance legal certainty for innovators’ and accelerate market access while enabling authorities to identify emerging risks. Proportionality measures, such as allowing microenterprises to fulfil ‘one of the most costly obligations... in a simplified manner,’ aim to preserve SME participation without lowering protection standards. In the US Executive Order, compliance is embedded in a broader agenda to ‘promote competition in AI and related technologies,’ targeting risks from concentrated control of key inputs and encouraging ‘new opportunities for small businesses and entrepreneurs.’ This market-facing approach is combined with sector-specific oversight, where agencies may require AI tools to ensure compliance in housing or financial markets, linking competition policy to targeted fairness interventions. The Russian framework situates compliance within a tightly managed institutional order. Participants in experimental legal regimes remain responsible for ‘за соблюдение прав субъектов персональных данных’ [compliance with the rights of personal data subjects] for the entire duration of a project and beyond. Organisational policy is positioned as a governance instrument to maintain

‘effective control’ over operations, and legal responsibility for AI system decisions is non-delegable, with accountability assigned to identifiable human or legal actors. These differences reveal that, while all three systems acknowledge the need for oversight, the EU integrates it into market harmonisation, the US aligns it with competitive growth, and Russia embeds it in state-centric responsibility. This divergence shapes the regulatory conditions under which RS development takes place, influencing participation incentives and the scope of permissible design choices. The following section examines how these regulatory and economic structures intersect with cultural and ideological factors, influencing how regulatory and economic priorities are articulated in relation to RS governance.

4.4 Culture: Do current AI policies operationalise fairness, transparency, adaptability, and trust for RS?

Economic factors influence AI development, while cultural context shapes how regulatory priorities and expectations around AI systems are articulated. AI represents a cultural as well as a technical shift for users and for the wider technological ecosystem (Crawford and Calo, 2016). Regulatory texts do not emerge in isolation. They embed assumptions about risk, responsibility, authority, and social purpose. The analytical framework introduced earlier in this chapter is applied here to examine how these assumptions are expressed in governance discourse across the EU, US, and Russia. Consolidated into four categories of Governance and Authority, Individual versus Collective Orientations, Risk and Security, and Values and Social Harmony, the framework supports a comparative reading of how cultural priorities shape legislative focus.

This section considers how these cultural orientations are articulated in policy language and how they intersect with the design, deployment, and societal effects of RS. Three outputs support the analysis: keyword frequency, sentiment scores, and co-occurrence patterns, which are used to identify the prominence of cultural terms, measure their evaluative framing, and trace how they cluster within governance discourse. Together, they provide a multi-layered account of how each jurisdiction constructs cultural priorities and how these priorities influence the governance of RS. Cultural values are not treated as background conditions but as active forces that shape how RS are legitimised, regulated, and embedded in everyday information flows.

Attention to cultural framing is necessary because the meaning of terms such as fairness, transparency, or trust differs across jurisdictions, with fairness ranging from distributive justice in EU discourse to non-discrimination in US policy, and transparency denoting procedural openness in the EU, institutional reporting in the US, and systemic reliability in Russia. RS are not only technical infrastructures but also sites where regulatory assumptions about value, authority, and responsibility intersect with information organisation. Regulatory language encodes assumptions about authority, community, and responsibility, and these assumptions influence how RS are expected to function in practice. For instance, in contexts where authority is emphasised, RS may be positioned as tools for institutional control, while in contexts that foreground individual rights, they may be framed as instruments of personal choice. Without a cultural lens, governance risks being treated as purely technical, overlooking the ways in which RS embed particular social priorities into the organisation of information

flows. Cultural analysis therefore complements technical and economic perspectives by clarifying how normative commitments shape the informational environments that RS produce.

Governance and Authority

The Governance and Authority category examines how regulation frames hierarchy, control, and institutional structure. In the EU corpus (96,644 words), high frequencies for authority (537), compliance (156), and control (75) indicate a strong discursive emphasis on structured oversight and supervisory mechanisms within the regulatory text. This configuration aligns with the EU’s broader approach of embedding standardisation and accountability into AI governance. Sentiment scores for this category show positivity at 0.088, negativity at 0.077, and neutrality at 0.835, indicating an overall predominantly neutral tone with occasional aspirational framing in relation to authority, compliance, and control.

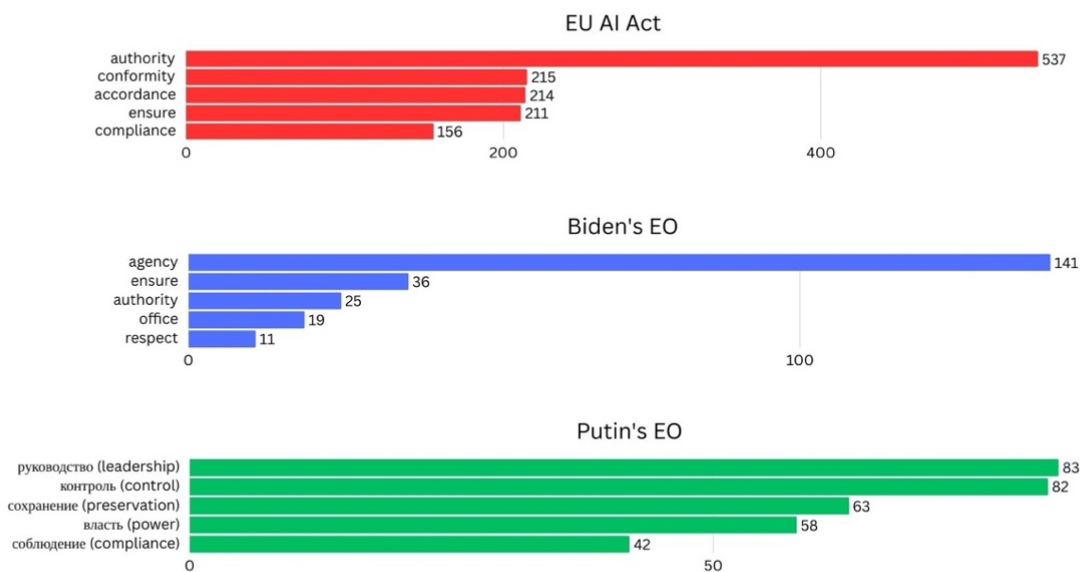


Figure 14 Top five most frequent terms in the Governance and Authority.

In the US corpus (21,088 words), authority (25) and control (4) appear far less frequently, indicating a reduced discursive emphasis on hierarchical oversight relative to the EU corpus. The sentiment profile here shows higher positivity (0.139) and similarly low negativity (0.038) compared to the EU, reflecting a discourse that frames authority as a supportive mechanism for enabling technological development and coordination rather than as a central organising principle. The Russian corpus (13,3845 words) contains frequent references to 'власть' [power] (58), 'контроль' [control] (82), and 'руководство' [leadership] (83), reflecting a governance style rooted in centralised state authority and executive coordination. Sentiment is overwhelmingly neutral at 0.995, with minimal positivity (0.003) and negligible negativity (0.002). This highly neutral tone indicates that references to authority are presented in a descriptive and procedural manner rather than framed through explicit evaluative language.

Individual versus Collective Orientations

The Individual versus Collective Orientations category captures how governance discourse balances personal autonomy with communal responsibility. In the EU corpus, references to individual focus dominate, with high frequencies for terms such as person (379) and personal (101), but these are paired with significant mentions of collective obligations such as obligation (218) and group (63). This combination reflects a regulatory discourse in which references to individual rights co-occur with obligations framed at a collective level, consistent with how privacy, autonomy, and responsibility are articulated in EU policy texts. Sentiment analysis for this category shows a predominantly neutral tone (0.894), with modest positivity (0.072) and low negativity (0.034),

indicating a cautious and procedural framing of the balance between individual and collective priorities.

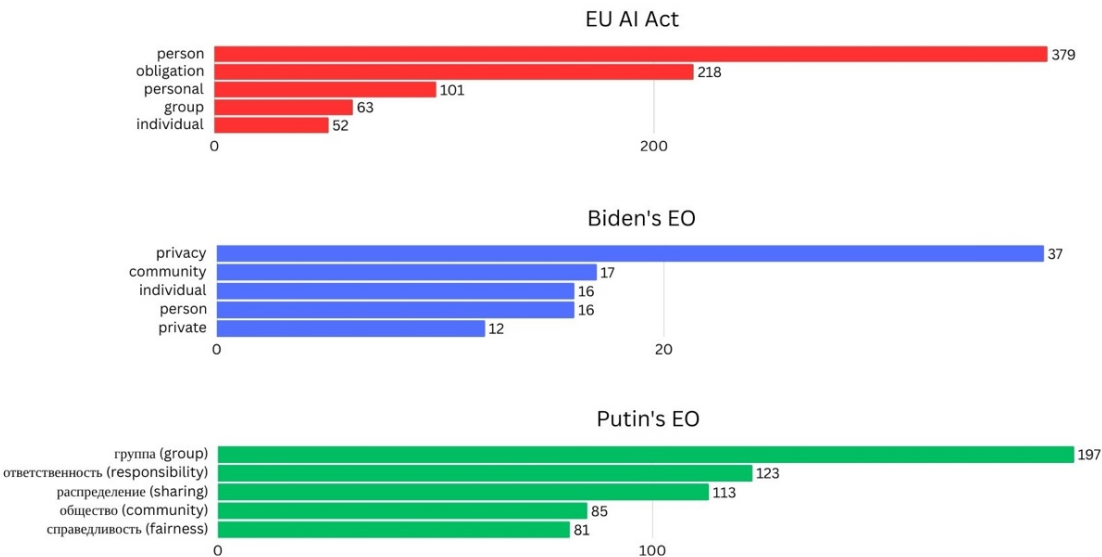


Figure 15 Top five most frequent terms in the Individual versus Collective Orientations category.

The US corpus shows a different balance, with a stronger focus on individual terms such as privacy (37), individual (16), and liberty (7), and fewer references to collective terms like community (17) and responsibility (5). This distribution suggests a discourse that foregrounds personal rights and freedoms, framing collective responsibilities as secondary or context specific. Sentiment in the US text is slightly more positive than in the EU (0.122) and comparably low in negativity (0.036), with a lower proportion of neutral language (0.842), reflecting a more optimistic and aspirational framing of autonomy and self-determination. In the Russian corpus, collective terms dominate the category, with particularly high frequencies for ‘группа’ [group] (197), ‘ответственность’ [responsibility] (123), and ‘распределение’ [distribution] (113). While individual terms such as ‘конфиденциальность’ [confidentiality] (70) and ‘автономия’ [autonomy] (16) appear, they are positioned alongside a much stronger emphasis on collective cohesion and role-based duties, aligning with a

governance model that places societal objectives above individual discretion. Sentiment analysis for Russia is almost entirely neutral (0.999), with negligible positivity (0.001) and no measurable negativity, signalling a formal, non-evaluative treatment of the relationship between individual and collective orientations.

Risk and Security

The Risk and Security category captures how governance discourse frames stability, protection, and preparedness in AI regulation. In the EU corpus, the most frequent terms are protection (146), safety (119), and security (54), reflecting a regulatory approach that places precaution and safeguarding at the centre of AI governance.

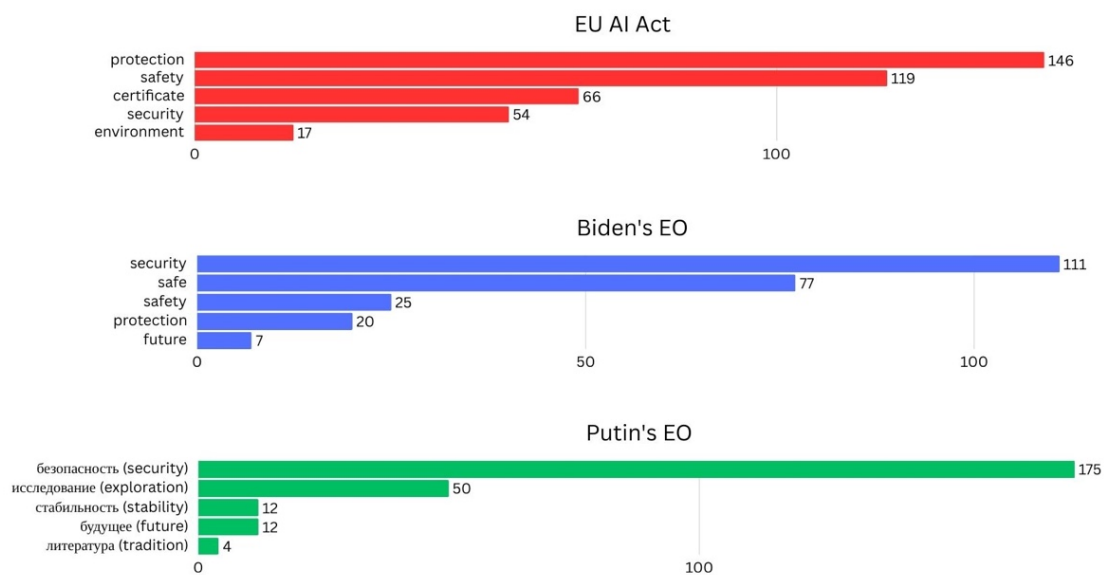


Figure 16 Top five most frequent terms in the Risk and Security category.

This emphasis is reinforced by sentiment scores showing moderate positivity (0.103) and low negativity (0.038), indicating that these measures are framed as necessary and beneficial to both system reliability and public trust. In the US corpus, terms relating to safety and security dominate, with security (111),

safe (77), and safety (25) counted separately rather than combined under a single stem in order to capture distinctions between institutional security, evaluative uses of 'safe,' and procedural references to 'safety.' The sentiment profile nevertheless differs markedly from the EU. The US registers the highest positivity of the three jurisdictions (0.231) with similarly low negativity (0.032), suggesting that safety and security are framed as enabling conditions for innovation and competitiveness rather than solely as protective safeguards. The Russian corpus is led by 'безопасность' [security] (175), followed at a distance by 'стабильность' [stability] (12) and 'будущее' [future] (12). Sentiment here is almost entirely neutral (0.997), with minimal positivity (0.003) and no measurable negativity, presenting security as a standard and unchallenged feature of state governance. These patterns suggest differences in how security-related terms are positioned within policy discourse across jurisdictions, with the EU emphasising precaution and trust, the US associating security with innovation-oriented language, and Russia presenting security in a largely descriptive and institutional register.

Values and Social Harmony

The Values and Social Harmony category in the EU corpus is defined by relatively low overall frequencies compared to other categories, with justice (16) as the most frequent fairness-related term and share and sharing together occurring 18 times as the leading terms in the universal good subcategory. Equality (7), deal (7), and fairness (5) also appear, indicating that the discourse embeds principles of equity and mutual benefit, albeit without the same intensity seen in governance or risk-related language. Sentiment analysis shows a mixed

but generally neutral-to-positive tone, with average positivity at 0.088, negativity at 0.077, and neutrality at 0.835. The range of compound scores suggests that while some occurrences frame fairness and equality in aspirational terms, others are linked to contexts of legal proportionality or dispute resolution, where evaluative language is more negative. This combination reflects a policy discourse in which references to fairness and social harmony are articulated alongside procedural and legal language, positioning these values within regulatory and compliance-oriented contexts.

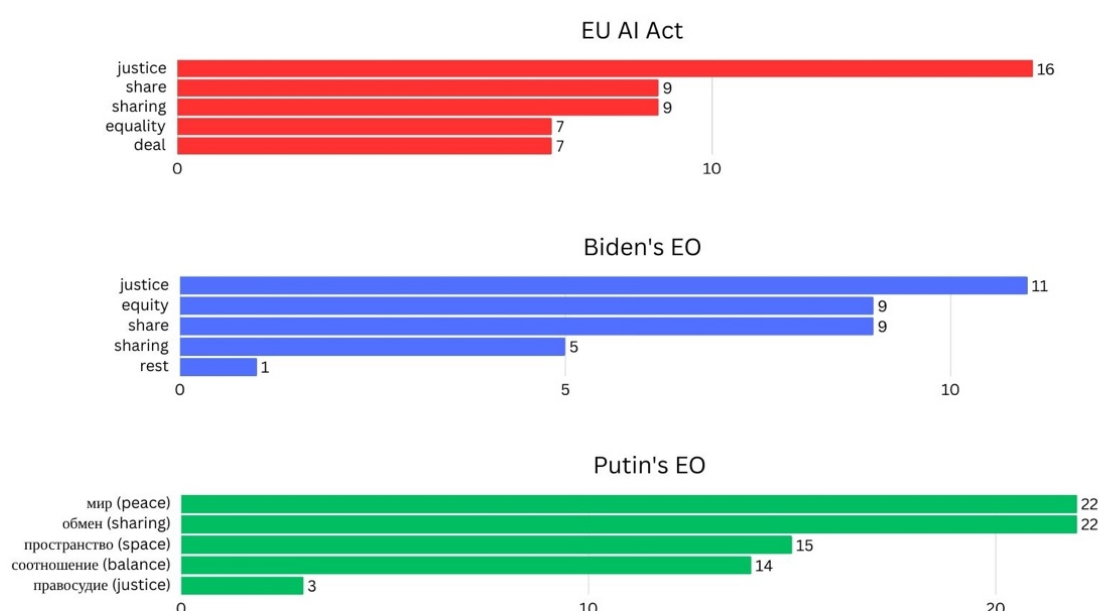


Figure 17 Top five most frequent terms in the Values and Social Harmony category.

In the US corpus, the category is anchored by fairness-related terms such as justice (11) and equity (9), with share (9) and sharing (5) representing universal good. The distribution suggests a relatively balanced attention to procedural fairness and communal benefit. Sentiment scores show positivity at 0.192, negativity at 0.044, and neutrality at 0.764, reflecting a more aspirational framing of fairness and inclusion than in the EU. The compound scores indicate that fairness and equality are often framed in aspirational or problem-solving contexts, producing a discourse that associates fairness and inclusion with

remedial and aspirational language, without specifying how such principles would be operationalised in RS governance. In the Russian corpus, the category features ‘мир’ [peace] (22), ‘обмен’ [sharing] (22), ‘пространство’ [space] (15), and ‘соотношение’ [proportion] (14) as the most frequent terms. This distribution reflects a discourse that frames social harmony in spatial, cooperative, and relational terms rather than through formal legal concepts. Sentiment is entirely neutral at 1.000, with no measurable positivity or negativity, presenting these values as fixed elements of the social order, aligning RS governance with the reinforcement of state-defined social cohesion.

The results of the cultural keyword, sentiment, and co-occurrence analysis show how each jurisdiction’s regulatory language reflects deeper ideological assumptions about authority, risk, individual responsibility, and societal purpose. These discursive patterns shape how AI systems such as RS are framed within regulatory texts, influencing how issues of civic participation, epistemic agency, and public accountability are articulated. In the EU, high-frequency co-occurrences such as ‘authority’ + ‘compliance’ (126) and ‘control’ + ‘responsibility’ (42) reveal a governance style that couples oversight with formal accountability mechanisms. This alignment is evident in provisions requiring that ‘market surveillance authorities shall verify the compliance with Article 60 as part of their supervisory role for the AI regulatory sandbox’ and in rules empowering the European Data Protection Supervisor to ‘impose administrative fines on Union institutions, bodies, offices and agencies’ with due regard to case-specific circumstances. Even in high-stakes contexts such as judicial decision support, where AI may be used ‘to assist a judicial authority in researching and

interpreting facts and the law,' the framing remains procedural, embedding technological deployment within pre-existing legal safeguards. This reflects high uncertainty avoidance and a cautious approach to technological risk, balancing individual rights with collective welfare through frequent references to community, responsibility, and equality alongside privacy and autonomy. In the Values and Social Harmony dimension, fairness appears more as a regulatory target than a cultural value, indicating a technocratic rather than inclusive narrative. In the United States, notable co-occurrences such as 'risk' + 'innovation' (58) and 'trust' + 'security' (41) highlight how regulatory discourse links the management of technological uncertainty to the promotion of opportunity, presenting safeguards not as constraints but as conditions that enable innovation and commercial expansion. High sentiment scores for autonomy and risk indicate confidence in self-regulation, and terms such as freedom, self, and trustworthy innovation frame governance as a facilitator of commercial dynamism. This approach is reflected in directives to develop 'standardized methodologies and tools for conducting and verifying the performance of sequence synthesis procurement screening... and processes for the reporting of concerning activity to enforcement entities' while celebrating AI's potential to make the world 'more prosperous, productive, innovative, and secure.' Such formulations balance technical risk management with an aspirational framing of technology as a driver of societal progress. While this enables decentralised civic participation, it does not guarantee equitable epistemic conditions, as RS may marginalise minority perspectives when engagement metrics drive optimisation. In Russia, dense co-occurrences such as 'контроль' [control] + 'ответственность' [responsibility] (87) and

‘безопасность’ [safety] + ‘власть’ [power] (53) signal the fusion of state authority, security, and duty. Accountability is defined as a condition in which an organisation ‘принимает на себя ответственность за воздействие на заинтересованные стороны, общество, экономику и окружающую среду... и не освобождается от ответственности за ошибочные решения, основанные, например, на результатах машинного обучения ИИ’ [takes responsibility for impacts on stakeholders, society, economy, and environment... and is not released from responsibility for erroneous decisions based, for example, on AI machine learning results], framing it as inalienable even when mediated by automation. Social values are articulated through terms such as ‘справедливость’ [fairness], defined as ‘подход, поведение или результат, основанные на уважении к установленным фактам, убеждениям и нормам и не определяемые предпочтением (фаворитизмом) или несправедливой дискриминацией’ [an approach, behaviour, or result based on respect for established facts, beliefs, and norms and not determined by preference (favouritism) or unfair discrimination]. Collective referents such as ‘общество’ [society], combined with minimal emphasis on individual choice, position AI development within a framework of national strategy and infrastructural governance. This framing positions RS within a discourse of state-led oversight and enforcement, articulating technological objectives in relation to centralised institutional priorities. These cultural framings were derived from keyword frequencies that identify salient concepts, sentiment scores that register their evaluative orientation, and co-occurrence patterns that show how they cluster, together revealing how regulatory logic shapes the treatment of RS.

4.5 Quality: Do current AI policies operationalise fairness, transparency, adaptability, and trust for RS?

Assessing quality in AI regulation requires examining how frameworks address fairness, transparency, adaptability, and trust. For RS, quality extends beyond output accuracy to include the capacity to support diverse perspectives, maintain user agency, and enable informed participation. The definition of a ‘good’ recommendation varies across contexts (Djeffal, Hitrova and Magrani, 2021): in commercial settings, relevance and engagement may be prioritised, while in civic contexts such priorities can diminish pluralism and accountability. Fairness concerns the equitable distribution of outputs across users and groups. Transparency relates to the extent to which users can understand how content is selected. Adaptability refers to responsiveness to changing contexts. Epistemic diversity denotes the system’s capacity to present a range of perspectives. Frameworks that neglect these criteria risk producing RS that are operationally efficient while leaving questions about their implications for informational conditions relevant to civic participation unresolved.

The governance frameworks of the EU, US, and Russia show limited engagement with RS-specific risks. Across all three, regulatory language remains general, and core issues such as bias, fairness, overfitting, and epistemic distortion are framed abstractly or omitted. This produces a gap in which RS are acknowledged as part of the AI ecosystem but are rarely regulated on their own terms. In the EU AI Act, bias appears mainly in connection with high-risk systems and seldom alongside accountability. RS are not classified as high risk, meaning that processes such as loss functions, which can prioritise

engagement and amplify inequities, remain unregulated. Transparency obligations are restricted to high-risk cases, leaving most RS exempt. In the US, Biden's Executive Order appeals to fairness and trustworthiness but offers no enforceable measures, relying on voluntary standards and sector-specific oversight. This decentralised model allows flexibility but leaves fairness and bias mitigation to platform discretion. In Russia, fairness is subordinated to state defined objectives, and transparency is framed as alignment with prescribed outputs. RS are positioned as tools for managing information flow, reinforcing dominant narratives rather than enabling pluralism. Overfitting, a technical risk in which recommendations narrow to past user behaviour, is unaddressed in all three frameworks. None provide requirements for generalisation, diversity, or iterative correction. Fairness is also weakly defined. In the EU AI Act, it is linked to item diversity in recommendations, which broadens exposure but does not address distributional equity. In the US, it is presented as an aspirational goal without operational metrics or enforcement. In Russia, it is defined through conformity to state normative values, positioning it as alignment rather than an independent standard for RS outputs. The cold start problem, which arises at system entry when limited user data forces reliance on profiling or third-party datasets, remains absent from regulatory guidance despite its potential to introduce structural bias during initial deployment. These omissions create conditions in which RS operate with minimal accountability while shaping what users see, consider relevant, and treat as credible. Broad terms such as 'trustworthy AI' and 'responsible innovation' signal an ambition toward ethical alignment, but lack procedural benchmarks, producing a compliance vacuum. For RS, which play a significant role in mediating access to information in some

contexts, this absence of targeted policy has important regulatory implications. Addressing it requires frameworks that explicitly define RS, set operational parameters, and embed fairness, transparency, and adaptability as mandatory design criteria.

Adaptability is essential for governing AI systems that evolve rapidly and operate across multiple domains. RS, in particular, present challenges that shift over time due to changing user behaviours, data inputs, and content types. A forward-facing governance model must be capable of tracking these changes and adjusting regulatory responses accordingly. Among the three governance frameworks analysed, only the EU AI Act contains language that signals such adaptability, but even this is limited in scope and does not apply directly to RS. The EU AI Act incorporates terms such as ‘continuous monitoring’ and ‘framework revision,’ indicating a regulatory logic that allows for updates based on technological developments. Provisions for post-deployment audits and documentation obligations for high-risk systems reinforce this model. However, RS are classified as minimal or no risk and are therefore excluded from most of these measures. This gap limits the extent to which the Act explicitly engages with potential long-term dynamics associated with RS, such as changes in recommendation patterns or cumulative effects of optimisation over time. Biden’s Executive Order in the United States emphasises ‘innovation,’ ‘partnership,’ and ‘voluntary standards’ but does not establish mechanisms for systematic evaluation or revision. It adopts a flexible stance that supports rapid market entry and sectoral autonomy but lacks adaptive infrastructure. As RS increasingly operate across domains, the absence of cross-sector monitoring provisions

creates a risk of regulatory lag. There are no requirements for reassessment once systems are deployed, nor is there language supporting longitudinal analysis of system behaviour. Russia's governance framework is static and ideologically anchored, containing no references to revision, audit, or post-deployment assessment and offering no accommodation for iterative improvement or adjustments in response to emerging harms or societal shifts. Corpus analysis shows that terms such as audit, monitoring, and revision appear frequently in the EU text, occasionally in the US text, and not at all in the Russian text. This pattern highlights differences in how adaptability-related concepts are articulated across the three governance frameworks.

The distribution of 'trust' and 'trustworthy' across the corpora exposes how governance texts construct confidence in AI. In the EU AI Act, 'trust' appears only twice, compared to nineteen instances of 'trustworthy.' This imbalance shows that the text stresses the assurance that systems are compliant and 'trustworthy' by regulatory standards, rather than addressing the relational dimension of how users come to place trust in RS through experience and interaction. Measures of user trust are inherently confounded by external factors such as interface design, habit formation, network effects, and exogenous media or political events, which limits their reliability as direct legislative targets. One occurrence links trust to accountability in high-risk contexts: '... retain public trust and ensure accountability and effective redress.' The other connects trust to the integrity of the information ecosystem, explicitly raising RS-relevant risks such as misinformation, manipulation, and synthetic content: '...significant impact on the integrity and trust in the information ecosystem, raising new risks of

misinformation and manipulation at scale...' Transparency is articulated primarily through procedural language, with limited specification of scope or application. The phrase 'without prejudice to other transparency obligations for deployers' leaves the scope and application of these obligations undefined: 'Compliance with the transparency obligations...should not be interpreted as indicating...lawfulness...and should be without prejudice to other transparency obligations for deployers...' In the US corpus, 'trust' appears three times (1.42 per 10,000 words) against 125 instances of 'trustworthy' (59.31 per 10,000 words), an even sharper imbalance given the smaller document size. One occurrence ties trust to human oversight: '...appropriate human review of AI decisions...to enhance trust and safety...' Here, trust is the outcome of procedural checks rather than a default property. Transparency is framed as an organisational responsibility: '...annual... reporting on how...risks from AI use cases... are managed...' A further example reinforces that transparency rests on regulated entities rather than the explainability of models: '...clarifying the responsibility of regulated entities...and...transparency of AI models...' In the Russian corpus, 'доверие' [trust] appears in relational contexts that anthropomorphise AI: 'Установление доверия между машинами и людьми работает аналогично...между людьми' [establishing trust between machines and people works similarly...between people]. Another frames trust as essential for acceptance of AI outputs: 'Одним из важных факторов такого доверия служит доверие к результатам функционирования систем искусственного интеллекта' [one of the important factors of such trust is trust in the results of the functioning of artificial intelligence systems]. Transparency is tied to controlling personal data systems, linking explainability to data traceability and

security: 'Эти элементы влияют на объяснимость, прозрачность, безопасность и конфиденциальность, особенно в случаях создания, контроля или обработки персональной идентифицируемой информации.'

[these elements affect explainability, transparency, security and confidentiality, especially in cases of creating, controlling or processing personally identifiable information]. Across all three jurisdictions, these patterns indicate that trust and transparency are discussed largely at an abstract or organisational level, with limited explicit connection to RS-specific quality considerations. The imbalance between 'trust' and 'trustworthy,' procedural rather than operational definitions of transparency, and anthropomorphisation of AI in Russian texts all indicate that confidence is constructed more as assertion or analogy than as a verifiable design outcome, reinforcing RS's marginal position within quality provisions.

4.6 Conclusion

Across the four analytical dimensions, this chapter has shown that the EU, US, and Russian AI policy frameworks address RS through broad AI provisions that overlook system-specific risks. In the Technology dimension, the absence of a clear definition leaves challenges such as cold start, bias, echo chambers, and overfitting unaddressed, with RS absorbed into generic AI categories. In the Economy dimension, the EU combines compliance with market integration, the US favours innovation with limited up-front regulation, and Russia concentrates development under state authority. Each model embeds an economic logic that shapes RS design priorities from engagement optimisation to ideological alignment without guaranteeing pluralism or accountability. In the Culture dimension, governance language reflects distinct normative orientations. The EU

favours procedural oversight, the US promotes decentralised opportunity, and Russia emphasises state-directed cohesion. These orientations affect how authority, rights, and obligations are balanced in RS governance. In the Quality dimension, none of the frameworks operationalise fairness, transparency, adaptability, or trust for RS. The imbalance between ‘trust’ and ‘trustworthy’ in EU and US texts, and the anthropomorphisation of AI in Russian discourse, construct confidence as an assertion or analogy rather than a measurable design property. These findings point to a misalignment between the language of AI policy and the RS-related epistemic and civic concerns discussed in this thesis. By embedding RS within general AI categories, current frameworks leave limited scope for addressing design practices that may be associated with narrower informational diversity, constrained user agency, or persistent bias. This marginal position is maintained through procedural rather than operational definitions of quality attributes and by economic and cultural logics that prioritise compliance, innovation, or ideological cohesion over targeted risk mitigation.

The next chapter examines how these policy framings circulate within RS discourse. It considers whether research on RS development and governance reproduces, reframes, or resists the definitional, economic, cultural, and quality framings identified here. This sets the stage for an analysis of semantic drift, showing how policymakers’ vocabulary enters technical writing and how this diffusion shapes the epistemic functions and civic role of RS.

Chapter 5. Recommender systems: Research Discourse

This chapter examines how academic writing on recommender systems (RS) research frames governance, agency, risk, and technical function through a lexical–discursive analysis of terminology used in computer science and engineering research. Drawing on peer-reviewed RS research in English and Russian, it analyses how terminological patterns influence the accessibility of knowledge about recommendation technologies for non-specialist audiences such as policymakers, journalists, and civic actors who rely on clear terminology to interpret and question system design. These patterns reflect both technical developments and institutional logics, indicating how academic terminology may condition the ways non-specialist actors interpret and question algorithmic systems, primarily through indirect and mediated channels rather than direct engagement with research discourse. The discussion continues the investigation of the civic empowerment gap introduced earlier. Chapter 3 showed that RS structure civic participation by reinforcing engagement-driven biases and linguistic silos. Chapter 4 demonstrated that regulatory responses prioritise procedural transparency over structural fairness, leaving epistemic exclusion unaddressed. Here, the focus shifts to discourse produced by RS researchers. It traces how academic writing embeds and propagates institutional assumptions through shifts in terminology and framing.

5.1 Methodology

Moving beyond the earlier categories of Technology, Economy, Culture, and Quality (Webster, 2014), the analysis uses four thematic clusters: governance and accountability, agency and autonomy, technical

instrumentalism, and risk and mitigation. These clusters were derived through an iterative review of term frequency lists, co-occurrence networks, and glossary definitions drawn from the dataset of English and Russian RS discourse. The process began with the categories used throughout this dissertation, which reflected broad domains of RS discourse. However, closer examination of the corpus showed that many high-frequency and central terms could not be meaningfully interpreted within those domains without obscuring their institutional or normative significance. Terms such as ‘accountability,’ ‘risk,’ or ‘autonomy’ were used across technical and policy contexts but carried distinct implications for how systems are justified or contested. Mapping directly onto the earlier categories would not capture these distinctions because many terms cross technical and policy contexts. While related to the earlier Technology, Economy, Culture, and Quality categories, the new clusters refine the analysis by providing clearer distinctions. Elements of Technology are reflected in technical instrumentalism, Economy overlaps with both governance and accountability and risk and mitigation, Culture intersects with agency and autonomy, and many terms previously grouped under Quality are now framed as part of risk and mitigation. Governance and accountability brings together terms relating to oversight, compliance, and institutional responsibility, recognising that these shape not only formal regulation but also the perceived legitimacy of RS. Agency and autonomy groups terms associated with user decision-making, system control, and the capacity to act without constraint. Technical instrumentalism isolates terms that frame RS primarily through optimisation metrics, efficiency measures, and performance indicators. Risk and mitigation aggregates terms concerned with uncertainty, harm, and remedial action. By

reframing the analysis around these clusters, the study interprets linguistic patterns in RS research in relation to potential epistemic effects. Each cluster provides a lens for examining how academic writing encodes assumptions about who holds authority, what forms of participation are possible, and which risks are deemed actionable.

In linguistic research, agency is typically analysed through grammatical structures that indicate who initiates an action and who is affected by it (Duranti, 2005; Ahearn, 2001; Dowty, 1991; Hopper & Thompson, 1980). This differs from political science, where agency refers to the capacity of individuals or groups to act independently within institutional constraints, and from computer science, where it often denotes the ability of software or systems to operate without direct human intervention. Clarifying these distinctions is necessary because the corpora contain texts influenced by all three domains, which may use the same term but with different implications. For example, the word ‘autonomy’ in a machine learning context often describes system performance without supervision, whereas in democratic discourse it refers to individual or collective self-determination. Aligning these uses is critical to evaluating how RS discourse either expands or narrows the interpretive space available to different publics. When a single term denotes grammatical agency, institutional capacity, and system independence, interpretive ambiguity is produced. Such ambiguity obscures whether RS discourse refers to human actors, institutional processes, or autonomous systems. The absence of distinction enables technical language to mask underlying political or institutional assumptions. As a result, terminological convergence reflects semantic flexibility that can be read as

aligning with institutional authority, allowing discourse to appear precise while shifting between technical, institutional, and political registers. This dynamic complicates challenges to RS outcomes, since the vocabulary blurs whether objections concern technical design, institutional practice, or political framing, making it harder to pinpoint what is being addressed.

A corpus of English-language research texts was compiled using the openalex.org platform, the successor to Microsoft Academic, widely used as an ‘open index of scholarly papers’ (Chawla, 2022; Färber et al., 2023; Velez-Estevez et al., 2023; Priem, Piwowar & Orr, 2022). The search applied the topic ID ‘RSs and Techniques,’ which covers articles, book chapters, preprints, editorials, and dissertations, and was divided into three five-year periods: 2010–2014 (1,574 open-access results), 2015–2019 (4,076 results), and 2020–2024 (8,801 results). The category is dominated by computer science and engineering outputs, with limited representation from the social sciences, meaning that the corpus reflects primarily technical perspectives on RS. This disciplinary concentration can also be understood as evidence of omission, insofar as RS operate across social, political, economic, and cultural domains, yet are institutionally classified within library and indexing infrastructures as a computer science object. In this chapter, that classificatory narrowing is treated analytically to examine how such institutional ordering shapes which concerns, actors, and forms of agency become visible in mainstream RS research discourse. This is evident from the distribution of host domains, where arXiv, ACM Digital Library, and IEEE account for about 33% of the corpus in 2015–2019 and about 51% in 2020–2024, with most of the remaining material also concentrated in technical

publisher platforms. Segmenting the data into five-year intervals follows common practice in computational social science and science mapping, enabling comparative and temporal trend analysis (Rotolo et al., 2017; Fortunato et al., 2018). File links were retrieved through the OpenAlex API and saved in a CSV file with their corresponding titles. PDF files were downloaded via a combination of direct HTTP requests and Selenium. After removing empty or error files, the corpus comprised 1,071 texts for en_10_14, 2,463 for en_15_19, and 4,392 for en_20_24. This reduction is consistent with Jones et al., who reported that ‘3 out of 4 links to a research text leads to changed content’ (Jones et al., 2016). Text extraction used Tesseract, an OCR tool widely adopted in academic research for its open-access distribution model (Smith, 2007; Pendlebury et al., 2019; Patel, Patel & Patel, 2012; Kay, 2007). Approximately 10% of additional files were excluded during OCR processing as unreadable, introducing a possible sampling bias by removing visually dense papers and potentially skewing the dataset toward more textually standardised publications. Exclusion of diagram-rich or industry-oriented outputs narrows the discursive range to those examples following norms for textually standardised formats. This constraint emphasises forms of reasoning that are institutionally validated through academic publishing norms, privileging analytic styles that lend themselves to textual analysis. Such skew does not undermine the validity of analysis but indicates that the most salient trends identified here are characteristic of conventional scholarly communication. The finding highlights how methodological filtering interacts with epistemic framing, foregrounding certain forms of discourse while rendering others less visible within the analysed corpus. To reduce this bias, excluded items were cross-checked and re-added where possible. The final corpus

supports the identification of long-term lexical shifts and institutional framing patterns in English-language RS research.

The Russian-language corpus was compiled using eLIBRARY.ru, the largest online repository of academic literature in Russian (Clarivate, 2025). The search used the keyword 'рекомендательные системы' [RS] and applied the same three five-year periods as the English dataset: 2010–2014 (99 open-access results), 2015–2019 (428 results), and 2020–2024 (1,017 results). As eLIBRARY does not provide direct download links and files are embedded in JavaScript, Selenium was used to automate PDF retrieval via document IDs, replicating manual access. The downloaded files had a failure rate similar to the English corpus: ru_10_14 contained 86 files (down from 99), ru_15_19 contained 371 (down from 428), and ru_20_24 contained 829 (down from 1,017). This lower loss rate may reflect differences in archival practices and publication infrastructures associated with some Russian research institutions, alongside the smaller scale of the dataset. The smaller scale of the dataset compared to the OpenAlex corpus also contributed to more consistent file recovery. All files were processed into text using Tesseract, following the same OCR pipeline as the English corpus. Together, the two corpora support comparative analysis of terminological shifts across distinct regulatory and linguistic contexts. Russian-language publications reflect not only disciplinary trends, but also institutional structures embedded in national indexing systems. State funding of research institutions and the classificatory practices of eLIBRARY reinforce alignment between discursive shifts and regulatory priorities. Metadata, keywords, and categorisation in this archive reproduce state-sanctioned frameworks that shape

how scholarship is organised and retrieved. In this environment, frequency patterns are conditioned by the institutional logics of knowledge production and funding rather than solely by disciplinary usage. The Russian corpus therefore demonstrates how infrastructural arrangements of academic publishing influence the circulation of terminology, with state-defined categories functioning as an additional layer of semantic framing. Comparative analysis accounts for this embedding, since differences across corpora extend beyond linguistic variation to include systemic conditions of discourse formation.

Both English and Russian corpora were pre-processed before analysis. English texts were lemmatised using WordNet, and Russian texts using RuWordNet, to reduce lexical variation and treat semantically equivalent word forms as single terms. All texts were converted to lowercase, stripped of punctuation and boilerplate content (e.g., journal mentions, dates), and filtered to remove stop words and words shorter than three characters. A predefined list of ethics- and governance-related terms was compiled from the Ethically Aligned Design First Edition Glossary (IEEE, 2017). This list served as both a thematic guide and the primary filter for empirical analysis. For each period (2010–14, 2015–19, 2020–24), word frequency and co-occurrence metrics were calculated exclusively for terms in this controlled vocabulary. Pairwise co-occurrences were extracted, with nodes representing terms and edge weights representing co-occurrence frequency, and visualisations generated in Cytoscape. Network centrality measures (degree, betweenness, closeness) were calculated to identify key terms. These metrics mapped term prominence and community affiliation but did not capture rhetorical positioning or sentiment. This combination

of frequency-based filtering, network analysis, and qualitative illustration supports a layered description of how shifts in RS research terminology relate to patterns relevant to epistemic agency and civic intelligibility. Custom metrics calculated percentage changes in term frequency and directional trends to complement the network analysis. Frequencies are presented in normalised form (per million or per ten thousand words) to account for differences in corpus size and enable cross-period comparison. Co-occurrence counts, however, are reported in raw numbers to preserve the scale of associative patterns, with normalised values used only in visualisations where explicitly indicated. Results were saved to structured CSV files, and co-occurrence heatmaps visualised intensity patterns across periods.

Terms from the Ethically Aligned Design First Edition Glossary (IEEE, 2017) were assigned to the four thematic clusters that structure the analysis in sections 5.2 to 5.5: governance and accountability, agency and autonomy, technical instrumentalism, and risk and mitigation. Within this set, terms that appeared frequently, demonstrated consistent use across periods, and clustered with related concepts in the network visualisations were included. Clustering was based on both semantic function and relational positioning within the discourse. Selection combined conceptual coherence with structural evidence from co-occurrence networks. Terms were grouped when they regularly occurred in the same communities over time and shared functional or institutional logics. For instance, ‘transparency,’ ‘audit’ and ‘accountability’ consistently co-occurred with regulatory language and institutional actors, supporting their placement in the governance cluster. Representative excerpts were then retrieved manually from

the cleaned corpus for each cluster. Selection focused on recurring collocations and surface-level agent–action formulations observable in the text, without analysing rhetorical positioning or interpretive intent. For instance, ‘ensuring transparency’ was contrasted with ‘transparency is ensured by government’ to illustrate differences in agency attribution and institutional responsibility. Comparable excerpt selection was carried out for the other clusters. For agency and autonomy, examples comparing active and passive constructions of user agency. For technical instrumentalism, excerpts highlighted abstract technical formulations without explicit human actors. For risk and mitigation, examples were selected to compare normative ethical framing with procedural mitigation language. These excerpts were not coded for sentiment or interpretation but served to anchor the structural findings in authentic language use. This combination of frequency-based filtering, network analysis, and qualitative illustration supports a layered account of how shifts in RS discourse shape epistemic agency and civic intelligibility.

Cluster analysis displays how terms gain meaning through association rather than isolated use. The governance cluster links transparency with regulatory language, turning a technical property into a marker of institutional legitimacy. The agency cluster shows autonomy appearing with both user-centred and machine-centred language, so the same term denotes human control and system independence. Such dual usage is not visible in frequency counts alone but becomes clear in relational patterns. Clusters, therefore, map the semantic environments of RS discourse, showing how vocabularies stabilise

assumptions about authority, participation, and risk. This makes concealed priorities in academic writing systematically identifiable.

This approach aligns with research examining how computer science discourse embeds normative positions within technical framing. Studies such as Hadar et al. show how design practices and developer mindsets shape the operationalisation of privacy (Hadar et al., 2018). Birhane and Birhane et al. demonstrate how machine learning research encodes specific values through framing choices, influencing how systems are justified and contested (Birhane et al., 2022; Birhane, 2021). Weidinger et al. highlight how the language of large-scale model research frames ethical and social risks in ways that may narrow the scope of perceived harms (Weidinger et al., 2021). Building on these insights, the following analysis examines how RS research frames oversight, compliance, and legitimacy, beginning with the governance and accountability cluster.

5.2 Governance & Accountability

Transparency, accountability, and auditability are recognised as core elements of democratic governance because they enable oversight, contestation, and legitimacy. In political theory and institutional practice, these concepts allow citizens to scrutinise decisions, question authority, and interpret how power is exercised through technical infrastructures. Within RS policy, they mediate the relationship between system behaviour and public intelligibility by providing procedural means for civic trust to be established and maintained. When these means are absent or restricted, users have limited capacity to examine the consequences of algorithmic systems or influence their normative

orientation. Trust in this context is not generated solely through technical implementation. It depends on institutional accountability, clear communication, transparency, and opportunities for user participation in governance (Knowles & Richards, 2021; Herrera-Poyatos et al., 2025; Cheong, 2024). These dimensions extend beyond technical performance or compliance metrics, functioning as institutional conditions for epistemic and political inclusion. In RS governance, transparency and accountability can operate both as normative aspirations and as operational practices. In discourse, their presence may indicate a commitment to democratic principles, while their framing or omission can reveal orientations towards either public inclusion or institutional control. In Responsible AI discourse, these terms frequently appear in ways that resemble performative assurances, where ethical considerations are referenced without accompanying detail on mechanisms for inclusion or redress. Their rhetorical role can outweigh their operational substance. References to ‘transparency’ and ‘accountability’ feature prominently in guidelines and frameworks, where they act as markers of procedural ethics. The extent to which these procedural ethics translate into meaningful epistemic agency depends on whether users can influence the principles guiding system design, rather than merely observe compliance. In practice, when accountability is defined only through documentation and procedural visibility, the scope for user intervention in shaping recommendation logic is minimal. This usage is effective within regulatory compliance regimes, where demonstrating adherence to prescribed norms carries greater weight than enabling participatory governance. In such contexts, these concepts operate as boundary objects that maintain trust among regulators, developers, and the public, but within legally predefined parameters. Economic incentives influence

how ethical concerns are addressed, with accountability discourse frequently framed in terms of compliance. This framing centres on documenting system behaviour and evidencing conformity with authorised procedures, without interrogating development practices that align compliance with economic objectives in ways that may constrain civic participation. This narrowing of accountability into compliance language is analytically relevant to the epistemic agency gap, as it shifts discourse from enabling critical interpretation towards validating pre-set operational norms. The prevalence of this framing suggests that researchers often respond to external legal frameworks, including the EU AI Act, rather than embedding fairness as an intrinsic design principle. Under such conditions, fairness becomes positioned as a compliance target rather than a normative commitment. The growing emphasis on accountability, shaped by regulatory trends, is evident in how privacy and transparency are treated as economic priorities (Viljoen, 2021). Transparency and accountability often appear as conditions of legitimacy, but their use tends to remain procedural rather than substantive. RS discourse follows this pattern by treating these terms as properties of systems or regulations without showing how participation is secured. This creates a gap between procedural legitimacy, marked by oversight language, and substantive legitimacy, grounded in negotiation with affected publics. The discourse privileges the former, casting transparency and accountability as signs of reliability rather than as practices open to contestation. In doing so, governance terminology sustains institutional authority while limiting the scope of civic inclusion.

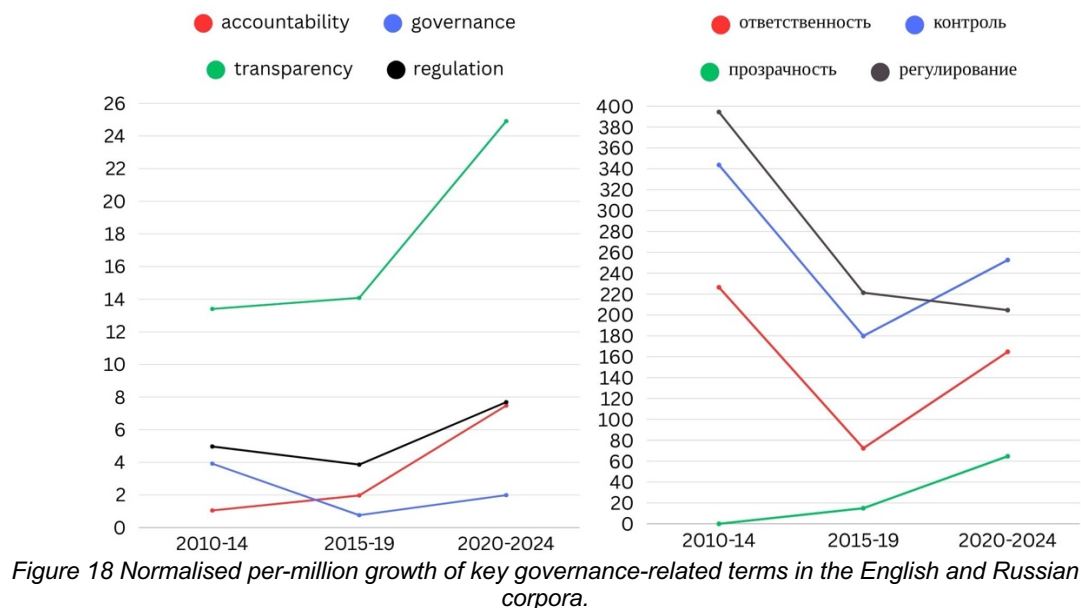
This shift from civic accountability to procedural simulation is a central concern in this chapter. It reflects what regulatory theorists describe as the displacement of moral reasoning by audit logic (Power, 1997; Strathern, 2000). In such contexts, transparency no longer enables scrutiny but confirms the existence of a compliant process. Audit trails substitute for contestation, and metrics of explainability replace public reasoning. These changes position democratic publics as recipients of procedural information rather than participants in shaping how algorithmic systems influence knowledge access, inclusion, and autonomy. Here, participation refers not to involvement in the technical development or governance of RS, but to broader democratic participation affected by RS outputs, including citizens' capacity to interpret information, engage in public discourse, and exercise epistemic agency. While participatory mechanisms in RS development are typically limited to expert, institutional, or stakeholder settings, the analysis focuses on how RS shape participation downstream, by structuring what publics can know, contest, or deliberate within democratic environments. This distinction matters because increased participation in compliance or oversight processes does not necessarily translate into enhanced democratic participation in the information environments shaped by RS. Legal frameworks such as the GDPR have contributed to increased research on compliance, particularly in relation to privacy, security, and explainability (Li, Yu & He, 2019; Goddard, 2017; Crockett, Goltz & Garratt, 2018; Chassang, 2017). This alignment allows RS developers to present systems as accountable without addressing how RS design affects civic participation or epistemic agency (Scott, 2018). In such an environment, technical refinement and legal adherence can take precedence over democratic

accountability (Novelli, Taddeo & Floridi, 2024; Doshi-Velez et al., 2019). The emphasis on compliance targets suggests a regulatory influence on research agendas, positioning them both as anticipatory alignment with legal requirements and as cost-management strategies (Vainionpää et al., 2024; Morgan, 2023; Lysaght, 2022; Calo, 2011). While state regulation applies external pressure, RS researchers also invest in internal governance mechanisms (Prunkl et al., 2021). As discussed in Chapter 4, this reflects a broader tension between legal accountability and institutional autonomy. A similar pattern appears in Russian-language RS research, though shaped by a different institutional logic. Here, governance and accountability are subordinated to national regulatory priorities. Transparency is framed less as civic intelligibility and more as a tool for state oversight. Audit and disclosure are framed less in relation to trust-building and more in relation to control, compliance, and insulation from external influence. The co-occurrence of technical and governance terms in the Russian corpus points to a discourse where RS are situated within institutional rather than user-centred contexts, as illustrated in the following examples. Discussions of algorithmic transparency or user data protection emphasise mitigating risks of misuse and external interference rather than supporting democratic accountability (Tsygankov, 2014). This model does not require censorship through algorithmic design, as legal frameworks themselves define the boundaries of permissible discourse. The result is a discourse of infrastructural compliance in which fairness, accountability, and transparency serve to legitimise institutional arrangements. The comparative drift identified here moves from participatory oversight to procedural signalling in the English corpus, and from normative ethics to strategic control in the Russian corpus. In both cases,

the terms remain visible but their semantic scope narrows. They become indicators of alignment, either with international regulatory templates or with national strategies of informational order. The patterns of linguistic drift observed can be read in relation to the institutional logics that structure RS development, indicating how discourse aligns with differing conditions for user interpretation and contestation.

Term trends

In the English corpus, transparency shows the highest normalised frequency, rising from 13.4 to 24.9 occurrences per million words. Accountability grows more modestly, from 1.0 to 7.5, while regulation increases from 5.0 to 7.7. Governance remains comparatively marginal, stabilising around 2.0. In the Russian corpus, 'регулирование' [regulation] and 'контроль' [control] hold the highest normalised frequencies, despite their overall decline from earlier peaks. 'Ответственность' [accountability] increases after 2019, reaching 164.7 per million words, while 'прозрачность' [transparency] shows later growth, climbing to 64.7 per million. The non-glossary term 'контроль' [control] is among the most central in the Russian dataset, frequently functioning in the collocational role that audit occupies in English texts. References to audit and its Russian equivalents were not part of the glossary-based frequency analysis but appear in the qualitative review of co-occurrence patterns.



In English texts, the rise of accountability and regulation indicates a discourse increasingly shaped by legal frameworks and process-oriented language. Accountability is typically framed as a procedural property of systems that is documented, demonstrable, and embedded through compliance measures. Transparency follows a similar trajectory but shifts from communicative openness to compliance visibility, often linked to explainability frameworks or regulatory requirements such as the EU AI Act and GDPR, rather than to user intelligibility or civic oversight. These patterns are analytically relevant to epistemic agency, as they indicate how the communicative space for questioning system behaviour is framed in RS research. In the English case, procedural–technical integration can leave oversight to standardised processes, limiting pathways for context-specific redress. In the Russian case, legal–technical operationalisation can further centralise interpretive authority within state institutions, reducing pluralism in governance narratives. In the Russian corpus, ‘контроль’ [control] and ‘регулирование’ [regulation] dominate governance-related vocabulary, reflecting a model in which oversight is enacted

through enforceable technical and legal constraints rather than civic mechanisms. ‘Прозрачность’ [transparency] occurs less frequently than in English texts and rarely signals user-facing openness, often replaced by ‘безопасность’ [safety] or ‘контроль’ [control]. The rise of ‘ответственность’ [accountability] and ‘регулирование’ [regulation] since 2020 aligns with the consolidation of national AI regulatory priorities (Saveliev & Zhurenkov, 2021; Papyshv & Yarime, 2023; Kharitonova, 2023). Whereas English-language governance discourse increasingly incorporates self-regulatory mechanisms such as industry standards, codes of conduct, and corporate social responsibility, Russian-language texts present RS as components of centralised infrastructural governance. In English, governance is often tied to anticipatory compliance and institutional signalling. In Russian, it functions less as an explicit lexical category and more as a structural assumption: platforms are expected to operate within state-defined parameters and are assessed on alignment with national strategy.

Co-occurrence patterns across periods reveal how the framing of governance and accountability terms has shifted. In the English corpus, ‘accountability’ consistently pairs with ‘transparency’ in all three periods, moving from low-frequency associations in 2010–2014 to becoming the most frequent link by 2020–2024. Early associations also include ‘law,’ ‘interaction,’ and ‘user,’ suggesting a conceptual connection between accountability and relational or rights-based contexts. By 2015–2019, ‘trust,’ ‘government,’ and ‘health’ enter cooccurrences, indicating an expansion toward institutional and sector-specific domains. In 2020–2024, the cluster around ‘accountability’ broadens to include ‘system,’ ‘research,’ and ‘algorithm,’ which corresponds to a greater association

with technical processes and evidentiary contexts in the co-occurrence patterns. ‘Governance’ shows a different trajectory: in earlier periods, it aligns with varied institutional and conceptual terms, but by 2020–2024 it is dominated by ‘data,’ ‘system,’ and ‘model,’ corresponding to a clustering around terms related to informational management.

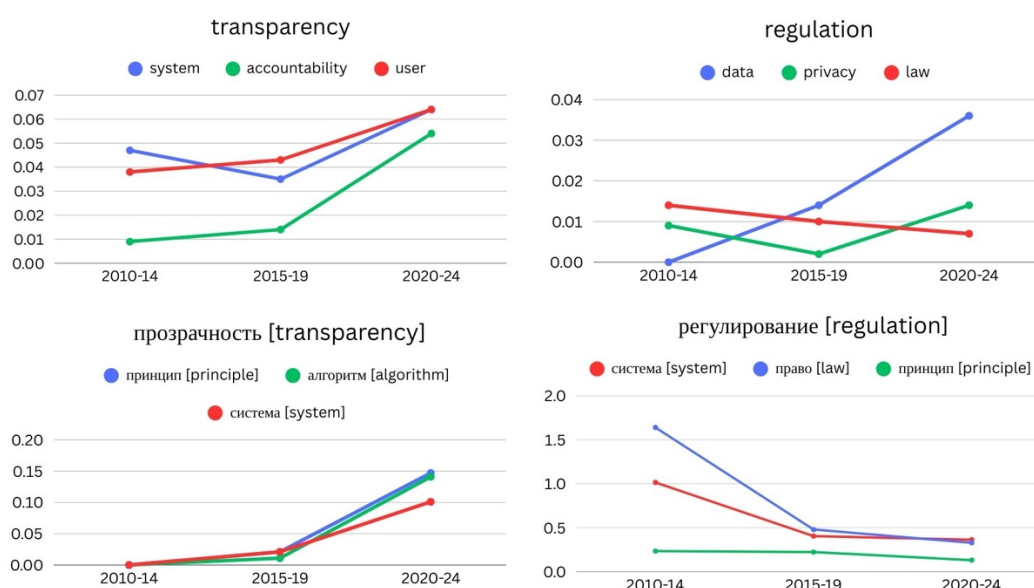


Figure 19 Normalised frequencies per 10k words of top co-occurring terms for transparency and regulation in the English corpus, and 'прозрачность' [transparency] and 'регулирование' [regulation] in the Russian corpus.

‘Regulation’ in 2010–2014 connects with general system-related vocabulary, while in later periods it acquires stronger links with ‘privacy,’ ‘law,’ and ‘compliance,’ reflecting increasing alignment with legal frameworks. ‘Transparency’ begins the decade linked to ‘system’ and ‘user,’ later gaining associations with ‘trust’ and ‘accountability,’ and by 2020–2024 expands to ‘model,’ ‘data,’ and ‘algorithm,’ embedding openness within technical architectures. In the Russian corpus, ‘ответственность’ [accountability] in 2010–2014 pairs with ‘право’ [law], ‘контроль’ [control], and ‘система’ [system], combining legal authority with supervisory functions. By 2015–2019, ‘право’ [law] remains central, joined by ‘принцип’ [principle], ‘обязанность’ [duty], and

‘безопасность’ [safety], which indicates a growing normative and protective framing. In 2020–2024, ‘право’ [law] still leads but co-occurs with ‘реализация’ [implementation], ‘риск’ [risk], and ‘данные’ [data], suggesting a shift toward operational and evaluative contexts.

Table 2 Raw frequencies of top co-occurring terms for transparency, accountability, governance, and regulation in English- and Russian-language corpora (2010–2024).

Term	Top co-occurrence	2010-14	2015-19	2020-24
English	corpus sizes (words)	6,639,763	13,207,710	28,620,542
transparency	system	31	46	183
	accountability	6	18	155
	user	25	57	183
accountability	system	0	2	31
	research	0	2	14
	government	1	3	5
governance	data	3	4	21
	system	0	1	8
	cultural	0	0	3
regulation	data	0	19	102
	privacy	6	3	40
	law	9	13	19
Russian	corpus sizes (words)	256,083	939,647	1,979,397
прозрачность [transparency]	принцип [principle]	0	2	29
	алгоритм [algorithm]	0	1	28
	система [system]	0	2	20
ответственность [accountability]	право [law]	17	11	51
	система [system]	6	5	27
	принцип [principle]	6	8	17
контроль [control]	система [system]	21	52	127
	мониторинг [monitor]	1	1	43
	данные [data]	0	6	34
регулирование [regulation]	система [system]	26	38	72
	право [law]	42	45	65
	принцип [principle]	6	21	26

‘Регулирование’ [regulation] in 2010–2014 centres on ‘система’ [system] and ‘право’ [law], with later periods adding ‘искусственный интеллект’ [artificial intelligence], ‘принцип’ [principle], and ‘безопасность’ [safety]. ‘Прозрачность’ [transparency] begins with limited connections, primarily ‘принцип’ [principle] and ‘алгоритм’ [algorithm], but in 2020–2024 expands to include ‘система’

[system], indicating a shift from sparse references to a more formalised role within technical design contexts. ‘Контроль’ [control], though not in the glossary, is consistently central and pairs with ‘система’ [system], ‘мониторинг’ [monitoring], and ‘данные’ [data] in the latest period, emphasising continuous operational oversight. These trajectories suggest that in English-language RS research, governance-related terms have become increasingly integrated with technical and informational processes, often through procedural linkages between accountability, transparency, and system-level concepts. In Russian-language RS research, the dominant associations remain anchored in legal authority and supervisory structures, with later additions reflecting implementation and risk management priorities. Across both contexts, the movement over time is toward institutional embedding within technical and legal infrastructures, but the pathways differ: English discourse moves from relational and rights-oriented associations to procedural–technical integration, while Russian discourse moves from legal–supervisory associations to legal–technical operationalisation.

Corpus examples

To illustrate how governance-related terms operate within academic RS discourse, this section presents representative excerpts from both English and Russian corpora. The examples are drawn from high-frequency co-occurrence contexts for the focus terms identified above. Each example reflects the most prominent co-occurring partner for the term within a given period, highlighting how meaning is shaped by grammatical construction, agency attribution, and institutional framing. In the English corpus for 2010–2014, accountability and

transparency appear together in contexts that connect them to law, policy-making, and civic participation:

It implies a set of common standards for action, which are defined by law and enforced in practice through procedures and while inclusive policy making enhances transparency, accountability and public participation and builds civic capacity, it also offers a way for governments to improve their policy performance by working with citizens, civil society organizations, businesses and other stakeholders to deliver concrete mechanisms for reliability, predictability and administration through law. Rule of law has been considered as one of the key dimensions that determine the quality and good governance of a country. (OECD, 2013)

The pairing with law and public participation reflects an early framing in which governance-related terms are linked to relational and rights-based contexts. By 2015–2019, the co-occurrence of accountability and transparency appears in more direct citizen–government relations:

In other words, the system aims at solving the previously mentioned government issues transparency, trust on the government, decision-making, and accountability in a cheap way as transparency there will be a connection between the citizens and the government, so the citizens can know everything happen in their government and in their city. (Al Helal & Mokhtar, 2018)

This framing positions the terms as communicative bridges between state and citizen, sustaining participatory connotations. In 2020–2024, the co-occurrence context shifts toward principles-based regulation:

Based on the latest regulations, businesses must develop their AI responsibly so that the AI system fulfils some vital principles, including transparency, fairness, accountability, privacy, and empathy. (Mansoorizadeh, Shahidi & Nazari, 2023)

Here, the terms are embedded in a principles list, signalling procedural integration into compliance frameworks rather than open-ended civic negotiation.

In the Russian corpus for 2010–2014, ‘система’ [system] and ‘право’ [law]

appear together in contexts that combine legal norms with broader normative references:

указанная международная система включает в себя как нормы международного права так и иные социальные нормы, не носящие юридически обязательного характера [the specified international system includes both norms of international law and other social norms not of a legally binding nature]. (Петров & Халафян, 2013)

This pairing reflects a legal–normative framing in which the system is positioned within an international and social order. In 2015–2019, ‘система’ [system] and ‘регулирование’ [regulation] occur in operational and administrative contexts:

система направлена на анализ и оценку качества принимаемых решений, увеличение числа новых студентов, совершенствование образовательных программ, регулирование и удержание производственных издержек, а также анализ факультативной работы [system is aimed at the analysis and evaluation of the quality of decisions made increasing the number of new students improving educational programmes regulating and controlling production costs as well as the analysis of optional work]. (Акимова et al., 2018)

The association of system and regulation with cost control and performance optimisation illustrates the technocratic emphasis observed in the middle period. In 2020–2024, ‘ответственность’ [accountability] co-occurs with ‘государство’ [government] in contexts defining responsibility for autonomous systems:

ответственность применения должно нести лицо, которое управляет робототехническим комплексом или программирует и отдает приказ на применение сас, но конкретные формы и методы контроля со стороны человека должны оставаться на усмотрение государств [accountability for application must be borne by the person who operates the robotic complex or programs and gives the order to use SAS but the specific forms and methods of control by a human must remain at the discretion of government]. (Иванов, Корженяк & Lapikhina, 2021)

This explicitly assigns legal–technical responsibility while allowing states to define the scope and method of human control, reflecting the operationalised legal framing of the latest period. Across both languages, the examples illustrate how the co-occurrence patterns identified in the term trends analysis appear in actual text. In the earlier English examples, governance-related terms are tied to participatory and relational contexts. Over time, they are repositioned within procedural and regulatory frameworks. In Russian examples, early contexts mix legal and broader social norms, then move toward technocratic operationalisation, and finally toward state-defined legal–technical responsibility. These patterns are consistent with the broader contrasts identified in the quantitative analysis, where English-language RS discourse is more frequently associated with procedural–technical integration, and Russian-language RS discourse with legal–technical operationalisation.

Comparative insights with IEEE Glossary

The IEEE Ethically Aligned Design glossary defines accountability as the ability to trace actions and assign responsibility, spanning legal, technical, and ethical domains. Across its formulations, it emphasises formal mechanisms such as documentation, attribution, and compliance with contractual or regulatory obligations. Transparency is framed as making information accessible and interpretable, from procedural disclosure under public law to a pro-ethical condition that enables principles such as accountability, safety, and informed consent (IEEE, 2017: p. 81). Regulation refers to rules issued by authoritative bodies, encompassing supervisory mechanisms and standards applied through legal, economic, and technical controls. In the English corpus, transparency most

often signals system compliance, paired with verbs such as ‘ensured’ or ‘verified’ through internal procedures, with no clear agent. In the Russian corpus, ‘прозрачность’ [transparency] appears infrequently and is often replaced by ‘контроль’ [control] or ‘безопасность’ [safety], which position oversight as enforcement or security provision. Accountability in English is linked to documentation, protocols, and compliance frameworks, reflecting IEEE’s procedural framing but omitting the civic responsiveness present in some domains. In Russian, ‘ответственность’ [accountability] is embedded in legal or state norms and frequently ‘определяется нормативно-правовой базой’ [defined by the regulatory framework], leaving no scope for negotiation. Regulation in English co-occurs with privacy, law, and compliance, embedding it in legal–technical frameworks. In Russian, ‘регулирование’ [regulation] is tied to system-level administration, risk mitigation, and state policy execution. Across all three terms, the corpus evidence shows a contraction from IEEE’s broader framing, which includes public accessibility, ethical justification, and trust, toward narrower, institutional outputs. In English-language RS research, this aligns with audit culture, where ethical concepts are reduced to documented process. In Russian-language discourse, the substitution of civic-oriented terms with control-oriented ones reflects a governance logic centred on legal sovereignty rather than participatory oversight. In both cases, affected publics appear to be marginalised, and the terms seem to function less as anchors for epistemic agency. These shifts in meaning set the stage for examining how such terms operate in context, which the next section addresses through close analysis of corpus examples.

5.3 Agency & Autonomy

Epistemic agency can be understood as the capacity of individuals to participate in the production, interpretation, and contestation of knowledge (Engel, 2013; Pohlhaus, 2020; Elgin, 2013). In democratic settings, it supports the ability to question institutional narratives, challenge prevailing framings, and navigate information environments with a degree of autonomy. This involves more than access to information, it extends to evaluating the credibility, intent, and potential implications of content. Defining autonomy as a capacity to be managed links it to system design rather than user action. In technical texts, autonomy is equated with machine independence, shifting agency away from users. When applied to civic settings, this framing obscures interpretive and participatory capacities of those subject to RS outputs. Autonomy becomes a property of architecture rather than a practice of choice, reducing room for contestation. This shift can be read as recasting a political concept as a managerial one, indicating how technical definitions may narrow understandings of democratic agency. Within AI ethics, epistemic agency is embedded in frameworks that emphasise human-in-the-loop (HitL) design and contestability, which seek to maintain openness to scrutiny, redirection, and rejection. As Coeckelbergh (2023) and Tufekci (2017) note, democratic legitimacy in AI deployment depends not only on formal commitments to fairness or explainability but on users' capacity to interpret, resist, or reconfigure system outputs. In the context of RS design, autonomy is not limited to interface settings or preference adjustments. It includes the ability to shape interactions in context, engage critically with presented material, and negotiate the terms under which algorithmic mediation occurs. Framed in this way, epistemic agency functions as

a civic capacity embedded in discourse rather than as a discrete technical feature. Where language fails to represent users as knowledgeable actors, it reinforces configurations that position them as passive recipients of recommendation logic. When agency is absent from the framing, users' interpretive actions are rendered invisible in both research and practice. This invisibility reduces the discursive recognition of users as co-producers of knowledge, which is analytically relevant to discussions of the civic empowerment gap. The gradual shift from participatory agency to behavioural optimisation is central to the present analysis. RS discourse often reframes agency through passive constructions and technical abstractions. The term 'user' is frequently employed as a proxy for commercial value, with behavioural data serving to refine model performance (Qin et al., 2020; Pi et al., 2019). This shift is significant because it redefines agency not as active participation in knowledge and decision-making but as behavioural input optimised for commercial objectives, which narrows the scope for democratic engagement and epistemic autonomy. In these framings, interpretive capacity is backgrounded, and the user is primarily defined as a data source aligned with optimisation, engagement, and retention targets. Agency becomes visible not through deliberation or informed choice, but as a statistical outcome measured by dwell time, click-through rates, or A/B testing results.

Technical refinement in RS development tends to serve economic objectives rather than epistemic empowerment or informational diversity. This creates a feedback loop in which system success is measured by sustained interaction, not by the extent to which users develop civic understanding or

critical capacity. As shown in Chapter 3, such optimisation is associated with patterns in which RS surface particular narratives, not because the systems are intentionally instilled with biases, or because these narratives are politically neutral or factually accurate, but because they generate measurable engagement. This dynamic shapes both the operational behaviour of systems and the criteria by which RS research defines effectiveness. Under these conditions, agency becomes associated with behavioural conformity: users are studied for the duration and frequency of their engagement rather than for the range of perspectives they encounter or the interpretive choices they make. Evaluation frameworks reinforce this tendency by assessing technical outputs against statistical benchmarks such as accuracy, precision, or recall, rather than epistemic goals such as understanding, fairness, or autonomy. Operational definitions of relevance or satisfaction are grounded in behavioural proxies, which displace deliberative or participatory measures. Where fairness is addressed, it is often reduced to statistical parity, leaving questions of structural inequity or redress outside the analytical frame. In such models, users' capacity to identify, interpret, or challenge systemic disparities is not a primary concern. Fairness becomes another optimisable parameter within an existing architecture, subject to the same performance logic that governs other system variables. These dynamics reflect a broader techno-solutionist narrative in which RS are positioned as tools for managing attention, increasing efficiency, or aligning content to predefined objectives. Within this framing, critical engagement is treated as inefficiency, and autonomy is perceived less as a normative priority than as a barrier to predictive accuracy. Such framings matter because they redefine autonomy in ways that favour system predictability over public capacity

to question or reject recommendations. In both commercial and regulatory contexts, this redefinition narrows the practical meaning of user choice to pre-configured options. Civil society actors, policymakers, and users are not absent because of insufficient technical expertise, but because the discourse presents system components as mathematically fixed and therefore insulated from public challenge. This orientation is visible in both the English and the Russian corpus, though with distinct emphases. In the English dataset, RS are characterised as adaptive systems oriented toward engagement and personalisation, with terminology such as user experience, interface optimisation, and feedback loops implying responsiveness without interrogating the purpose or consequences of that responsiveness. In the Russian corpus, RS are more directly framed as infrastructures for managing attention and securing alignment. Here, RS are more often framed in relation to strategic and administrative objectives rather than commercial ones, with emphasis on alignment, stability, and governance priorities (Makhortykh & Bastian, 2022; Erbsen & Pöldre, 2023; Kuznetsova & Makhortykh, 2023). Chapter 3 illustrated this in the case of Dzen, where RS-mediated exposure patterns were structured to reinforce pro-government narratives and align user experience with state-defined priorities.

This does not imply that Russian systems lack variation. Dzen's architecture, for example, displayed lexical diversity, though this operated entirely within parameters defined by state regulation. Algorithmic censorship was unnecessary because the legal framework had already delimited the scope of permissible discourse. Within this environment, user agency was exercised only in navigating a pre-structured narrative space. The design imperative was

not to maximise autonomy or deliberation, but to manage exposure in line with authorised priorities. Under such conditions, RS can be interpreted as being framed in ways that prioritise managed exposure over the expansion of civic capacities. Taken together, these patterns point to a reduced discursive emphasis on epistemic agency. Users are less visible in the language of RS research and less frequently positioned as central to design priorities. Users are both underrepresented in the language of RS research and excluded from its design priorities. The dominant focus on predictive modelling, performance benchmarks, and optimisation leaves limited space for dissent, reconfiguration, or user-driven modification. In the English corpus, these constraints emerge primarily from commercial imperatives, while in the Russian corpus, from regulatory alignment. Across both contexts, autonomy is reframed less as the capacity to act and more as the capacity to be managed. The following term trends analysis examines how these dynamics are reflected in the shifting vocabulary of agency-related concepts across the two corpora of RS-related scientific discourse.

Term trends

This section examines three IEEE glossary terms relevant to epistemic agency in RS discourse: agency, autonomous, and autonomy. In the Russian corpus, the absence of a direct equivalent for agency is addressed through the use of ‘актор’ [actor], which reflects the role-oriented framing found in the data. Centrality metrics (degree and betweenness) were applied to identify the relational prominence of each term, and co-occurrence patterns were analysed to determine their proximity to regulatory, behavioural, or ethical concepts. In the

English corpus, autonomous shows the highest normalised frequency across all periods, ranging from 9.0 to 10.4 occurrences per million words, with strong degree centrality (0.239) and notable betweenness (0.012).

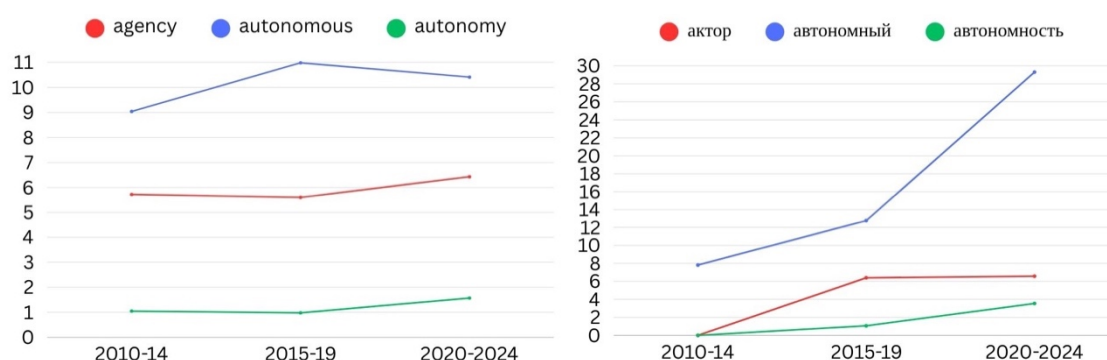


Figure 20 Normalised per-million growth of key Agency & Autonomy related terms in the English and Russian corpus.

Agency also increases modestly, from 5.7 to 6.4 occurrences per million words, with the highest degree centrality in this set (0.253) and moderate betweenness (0.010). Autonomy grows slightly, from 1.0 to 1.6 occurrences per million words, with lower degree (0.150) and betweenness (0.011) scores, and appears most often in functional contexts rather than in relation to deliberative capacity. In the Russian corpus, 'автономный' [autonomous] dominates, increasing from 7.8 to 29.3 occurrences per million words, with the highest degree (0.164) and betweenness (0.015) in this set. 'Актор' [actor] emerges after 2015 and stabilises around 6.5 occurrences per million words, with moderate centrality values (degree 0.116, betweenness 0.010). 'Автономность' [autonomy] remains infrequent, reaching 3.5 occurrences per million words in the most recent period, with low degree (0.066) and betweenness (0.003). The distribution shows a greater prominence of system-level autonomy terms in the Russian RS corpus relative to terms associated with participatory agency. Across both corpora, terms associated with system autonomy show the most

pronounced growth in both frequency and centrality. Terms denoting agency increase in English but remain limited in Russian, where *актор* [actor] appears in contexts that are more frequently associated with institutional or structural roles than with individual capacity. Autonomy in both contexts is less central and more narrowly framed. The following co-occurrence analysis examines how these terms are positioned within broader lexical networks in each corpus.

Co-occurrence analysis shows distinct patterns between the corpora. In the English dataset, agency increasingly clusters with research (rising from 9 to 47 co-occurrences), user (from 5 to 38), and government (from 6 to 19). Autonomous is most frequently linked to system (25 to 130), agent (47 to 100), and model (1 to 21). Autonomy co-occurs mainly with user (1 to 19), agent (2 to 5), and system (2 to 3). Although autonomy-related terms appear slightly more often over time, their collocational environment remains anchored in system functionality and institutional actors rather than in concepts of participatory choice.



Figure 21 Top co-occurring terms for agency, autonomous, and autonomy.

In the Russian dataset, co-occurrence levels are low. ‘Актор’ [actor] appears only sporadically and connects to terms such as ‘модель’ [model], ‘принцип’ [principle], ‘согласие’ [consent], ‘контроль’ [control], and ‘этика’ [ethics]. ‘Автономный’ [autonomous] links most often to ‘система’ [system], with occasional ties to ‘обучение’ [training] and ‘культура’ [culture]. ‘Автономность’ [autonomy] occurs infrequently and primarily with ‘система’ [system] and ‘безопасность’ [safety]. These results point to the near absence of agency and autonomy as descriptors of epistemic choice in Russian RS discourse, both conceptually and culturally, even in research contexts. Across both corpora, the growth of autonomy-related vocabulary does not translate into a shift toward user-led or participatory autonomy. Although Figure 19 shows increasing co-occurrence between autonomy and user in the English corpus, these references remain limited in scale and are embedded within a broader discursive structure in which autonomy is primarily articulated as a property of systems, models, and agents rather than as a capacity for user self-determination. The dominance identified here is therefore interpretive rather than purely quantitative: user autonomy appears more frequently, but system autonomy continues to organise how autonomy is defined, operationalised, and evaluated in RS research. In English, a modest increase in agency and autonomy is accompanied by clustering with technical and institutional terms, where autonomy refers not to complete independence but to the capacity of RS to operate without continuous human intervention. In Russian, low frequency and limited co-occurrence reinforce a framing in which agency is peripheral, and autonomy is treated as a property of systems rather than of individuals.

Corpus examples

Examples from the English corpus show that agency in relation to the user is often discussed within design and evaluation contexts but framed through technical affordances rather than participatory decision-making. In 2010–14, user agency is described as a potential outcome of interactive recommendation models, where conversational interfaces could create ‘a more concrete impression of agency’ through user contributions, yet the emphasis remains on benchmarking system success. In 2015–19, agency appears in studies noting the absence of ‘active control over the selection of stories,’ where randomised content insertion is interpreted as both a safeguard against over personalisation and a possible source of user dissatisfaction. By 2020–24, the term is tied to human-centred design principles, advocating that recommendations should enhance user agency without replacing it, while also stressing that transparency in system rationale is necessary to maintain trust. These co-occurrence findings reinforce the comparative picture: English discourse centres on system-responsive models that acknowledge but limit user input, whereas Russian discourse largely omits individual agency altogether. At the level of discourse, agency in the former is framed procedurally, while in the latter it is far less visible within RS-related contexts. In the Russian corpus, ‘актор’ [actor] is applied in abstract institutional or economic contexts, far removed from RS interaction. It is situated in subject domains such as economics or transport safety that are unrelated to information recommendation. ‘Автономность’ [autonomy] appears in discussions of transport safety and autonomous vehicles rather than epistemic choice in information environments. These examples reinforce the quantitative findings: in English RS discourse, agency is present but often operationalised

through system-centred mechanisms, whereas in Russian, its equivalents are conceptually and thematically peripheral to RS design and evaluation.

Comparative insights with IEEE Glossary

The IEEE glossary defines agency across disciplines as the capacity to decide and act, with computational and engineering domains framing it as a property of systems capable of autonomous decision-making, reactivity, proactiveness, and social interaction. In legal and ethical contexts, agency involves acting on behalf of another, or prioritising the well-being of others, with the capacity to exert influence. Autonomous is described as full independence in decision-making and operation, though engineering and legal definitions recognise that such independence may be partial or conditional. Autonomy is defined as the ability to govern oneself, encompassing the formation of goals and the execution of actions free from external constraint, with multiple forms recognised in computational literature, from environmental to norm-based autonomy. The corpus analysis shows that English RS discourse reflects a narrowed interpretation of these concepts, emphasising functional autonomy of systems and operationalised measures of agency, while largely excluding the ethical and participatory dimensions present in IEEE's definitions. In Russian RS discourse, 'актор' [actor] and 'автономность' [autonomy] appear in domains such as economics or transport safety, where they are detached from information recommendation contexts, indicating a limited integration of epistemic agency into RS-related conceptual frameworks. This absence reflects not only the lack of direct lexical equivalents but also the exclusion of these concepts from the thematic and methodological core of RS research in the Russian corpus. This

divergence suggests that, in both corpora, the breadth of meanings captured in the IEEE glossary is significantly reduced, with RS research privileging technical independence and system-level control over the fuller civic and ethical scope of agency and autonomy. These patterns, when set against the broader definitions in the IEEE glossary, highlight how RS discourse selectively adopts technical aspects of agency and autonomy while sidelining their ethical and participatory dimensions, which are central to preserving users' ability to contest, reinterpret, and influence recommendation processes.

5.4 Technical Instrumentalism

Instrumentalism refers to the use of technical language to strip social, political, and ethical context from technological development. In RS discourse, instrumentalism operates linguistically and structurally: terms such as algorithm, model, technology, and computation are used to describe system processes while concealing the normative assumptions embedded within them. This framing positions RS as optimisable artefacts governed by internal logic rather than contested systems shaped by human priorities. In this sense, instrumentalism can be read as aligning with a form of epistemic closure in RS development, where the scope of critique is narrowed and political choices are framed as engineering solutions. Disciplinary norms in computational and engineering research reinforce this orientation, privileging precision, optimisation, and reproducibility over civic or ethical considerations. The shift from conceptual discussions of AI to the concentrated language of algorithms, models, and computational performance reflects a consolidation of discourse around system internals. Where AI once encompassed societal stakes,

regulatory concerns, and public imagination, model now functions as a boundary term, localising responsibility and isolating technical performance from wider consequences (Helm et al., 2023; Krafft et al., 2019). This trend is evident in both English and Russian corpora. In English-language RS discourse, emphasis on algorithmic optimisation, architectural design, and computational efficiency has displaced questions of fairness, inclusion, and civic participation. As discussed in Chapter 3, politically consequential outputs such as demographic targeting or narrative reinforcement are not framed as social effects but as properties of algorithmic generalisation or computational constraints. In Russian-language RS research, the same technical vocabulary describes systems configured for strategic coherence, with models evaluated primarily by predictive accuracy and efficiency, rather than by reference to their potential role in shaping narratives or aligning with broader policy priorities.

The term algorithm appears in both corpora as a foundational component of RS design. It is generally discussed in terms of optimisation strategies, complexity, and implementation, which positions it as a neutral technical object. This presentation does not explicitly address the fact that algorithmic configurations reflect specific priorities, such as which patterns in the data are amplified or suppressed. The absence of discussion on how these priorities are set removes consideration of the political and cultural contexts in which algorithms operate. The term model functions as the primary analytical unit in RS research. In technical descriptions, models are treated as self-contained systems whose performance can be measured and improved without reference to the environments they shape. This perspective isolates the act of modelling from

questions about the origin, representativeness, or purpose of the data used. It also limits discussion of how model outcomes interact with social or institutional structures beyond the RS itself. References to technology often carry an implicit association with progress and innovation. In RS discourse, this association is rarely interrogated. Technology is framed as an inevitable and beneficial force, which masks the ways in which its adoption is mediated by institutional agendas and resource allocation. This framing limits discussion of who benefits from technological change and how such benefits are distributed. Computation is primarily presented as a measure of capacity and efficiency. It is described in terms of processing speed, scalability, or optimisation potential, without addressing the implications of what is being computed. This framing positions computation as an end in itself, discouraging scrutiny of whether computational processes support goals such as pluralism, accountability, or inclusivity, a tendency reinforced by the self-referential character of computer science as a discipline that often studies its own methods. These linguistic patterns reinforce a closed technical frame in which RS development is assessed by internal performance measures rather than by its role in shaping knowledge environments. In both the English and the Russian corpus, the terms function as reference points for technical capability while leaving questions of societal impact underexamined. The result is a discourse in which the technical properties of RS are foregrounded, and their civic dimensions are peripheral. Optimisation often functions as the central organising principle of RS discourse. Terms such as algorithm and model reinforce a vocabulary in which technical refinement stands as a measure of progress. This focus sidelines alternative framings that might highlight distributive or cultural concerns. The effect is not erasure but

marginalisation, as social and political issues are pushed to the edges of discussion. Optimisation discourse thus establishes technical priorities as the default framework, shaping an environment where wider epistemic or civic stakes appear secondary. Within this framing, technical language becomes the primary register of authority, narrowing the range of issues that are foregrounded in debate.

Term trends

This narrowing of meaning is directly reflected in the normalised frequency and centrality patterns of these terms, showing how technical framings consolidate around system-internal metrics. In the English corpus, ‘model’ holds the highest degree (0.828) and betweenness (0.032) scores in this set, while ‘algorithm’ also shows strong connectivity, with degree (0.705) and betweenness (0.016). ‘Technology’ and ‘computation’ appear less often, but ‘computation’ has comparatively high betweenness (0.021), suggesting a bridging role in technical discourse despite its lower overall frequency. In the Russian corpus, ‘модель’ [model] and ‘алгоритм’ [algorithm] dominate, while ‘реализация’ [implementation] and ‘технология’ [technology] show smaller but steady presence. ‘Реализация’ has the highest betweenness (0.052) in the Russian set, indicating its connective role across technical subfields. Across both corpora, the consolidation of centrality scores points to a discourse increasingly anchored in system internals and operational optimisation, with limited integration of social or ethical considerations.

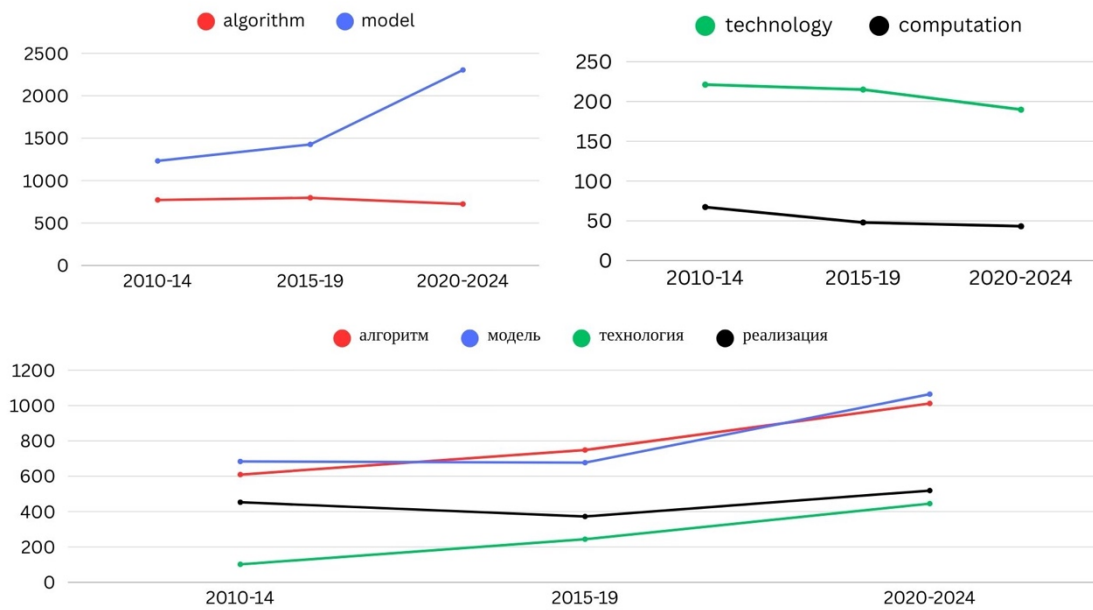


Figure 22 Normalised per-million growth of key Technical terms in the English and Russian corpus.

Co-occurrence patterns reinforce the concentration of these terms within technical frames. In the English corpus, algorithm most frequently co-occurs with user (rising from 1,291 to 4,324 occurrences), system (547 to 2,548), and model (410 to 2,416). The term user is analysed here because it functions as the dominant and most generalised stakeholder category in RS research discourse. Alternative terms such as citizen, data subject, or rights holder were considered, but occur far less frequently and are typically confined to legal, policy, or governance-oriented texts rather than to core RS research publications. For consistency across clusters, the analysis focuses on the top three co-occurring stakeholder terms by frequency within each period, with the relative absence of civic or rights-based categories treated as an empirical finding rather than as term exclusion. Model is strongly associated with user (3,137 to 14,967), data (769 to 6,971), and training (199 to 6,177), indicating its anchoring in performance-oriented contexts. Technology links primarily to information (569 to 1,861), system (336 to 1,001), and research (129 to 533), suggesting its role as

a broad infrastructural term rather than as a locus for social or ethical debate. Computation co-occurs with user (150 to 257), algorithm (55 to 167), and system (32 to 148), reinforcing its framing as a capability metric. In the Russian corpus, алгоритм [algorithm] clusters with система [system] (35 to 475), обучение [training] (3 to 330), and данные [data] (26 to 313), while модель [model] is most often linked to система [system] (30 to 360), данные [data] (27 to 349), and обучение [training] (18 to 336). Технология [technology] connects to система [system] (11 to 217), обучение [training] (1 to 105), and данные [data] (3 to 87). Реализация [implementation] is associated with система [system] (21 to 264), алгоритм [algorithm] (9 to 102), принцип [principle] (8 to 75), and политика [policy] (6 to 44).

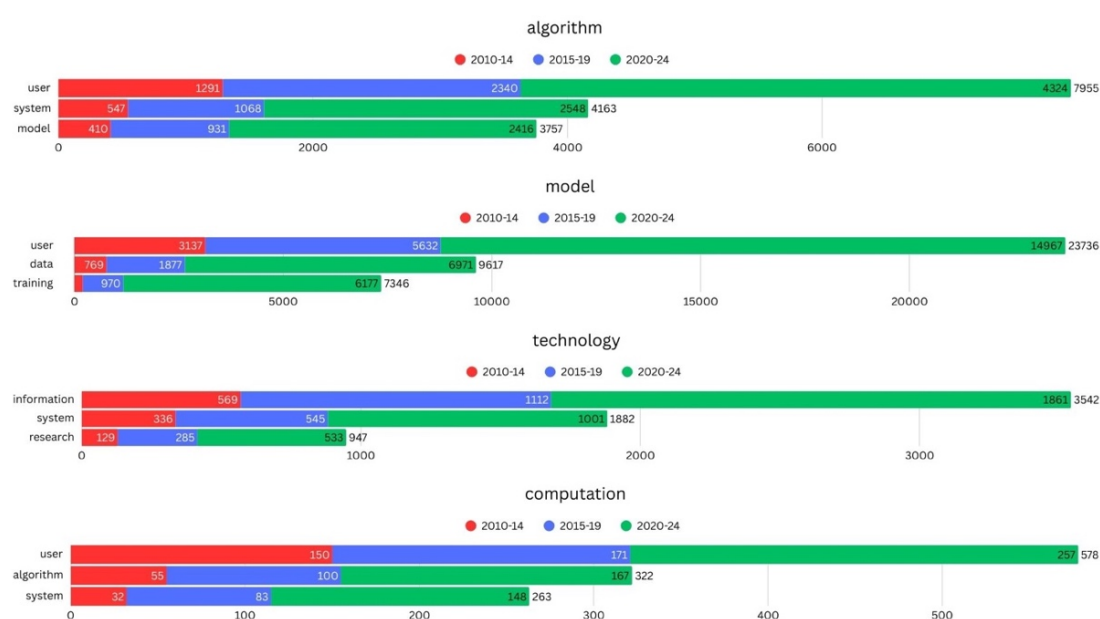


Figure 23 Top co-occurring terms for algorithm, model, technology, and computation.

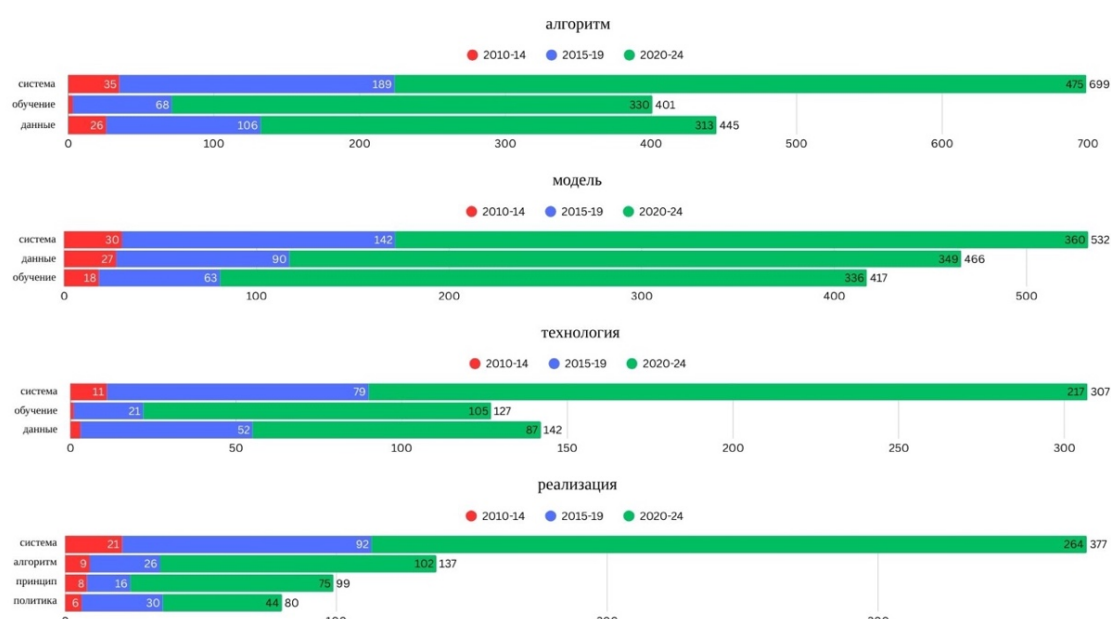


Figure 24 Top co-occurring terms for 'алгоритм' [algorithm], 'модель' [model], 'технология' [technology], and 'реализация' [computation].

Across both corpora, these co-occurrence patterns keep technical vocabulary embedded in closed operational loops, linking primarily to other system-internal concepts. This reinforces a discursive structure in which technical refinement is foregrounded while societal implications remain peripheral.

Corpus examples

Examples from the English corpus show that the terms model, algorithm, technology, and computation are embedded in highly technical contexts, with social or ethical implications remaining implicit or absent. References to model and data in 2010–14 caution against 'simply relating our data to different information-seeking models in order to identify a model or models which fitted the data best,' noting that this could be 'regarded as forcing as opposed to emergence in grounded theory terms.' By 2015–19, the term is situated in optimisation narratives, as in 'the output of the CNN model providing information

about unexpressed customer preferences has been seamlessly integrated with interaction data rating matrix in the proposed item-based recommendation algorithm.' In 2020–24, discussion turns to performance constraints, where 'the personal data on each client is generally limited, which can introduce biases and compromise model performance.' Mentions of algorithm and user move from early descriptions of 'a graph-based iteration algorithm to discover users' interested topics' in 2010–14, through rule-based 'pre-recommendation for internet users to browse news' in 2015–19, to contemporary evaluations where 'we use a particular user as test data and the remaining ones as train data...and compute predictions for all items for the user by running the considered algorithm on the train set.' In the Russian corpus, алгоритм [algorithm] and модель [model] appear in paired technical descriptions: 'все параметры модели, используя алгоритм, затем, используя полученные параметры, общий рейтинг неизвестного объекта предсказывается как наиболее вероятное значение' [all parameters of the model are applied using the algorithm, then using the obtained parameters, the overall rating of an unknown object is predicted as the most probable value], followed by applications to multi-criteria evaluations. In 2015–19, the framing is tied to machine learning sensitivity: 'в зависимости от встроенного алгоритма машинного обучения такие системы могут быть чувствительны к обучению' [depending on the built-in machine learning algorithm, such systems can be sensitive to training]. By 2020–24, algorithmic development is explicitly linked to socio-cultural contexts: 'введение инновационных алгоритмов решения конкретных проблем, связанных с многонациональным управлением, предполагает создание системы, специально нацеленной на решение трудностей, возникающих из-за

разнообразия культур и национальных характеристик' [the introduction of innovative algorithms for solving specific problems related to multinational governance implies the creation of a system specifically aimed at addressing difficulties arising from cultural and national diversity]. Even here, cultural and national diversity is instrumentalised as a technical design variable rather than as a space for participatory negotiation, illustrating how civic or deliberative framing remains marginal within the Russian RS discourse. Across both corpora, these examples illustrate RS discourse that is predominantly organised around procedural and performance-oriented framings, with broader epistemic and civic considerations appearing less frequently.

Comparative insights with IEEE Glossary

The IEEE glossary defines algorithm as a clearly specified sequence of rules for solving a problem or conducting a series of computations, with precision, termination, and reproducibility as core criteria. Model is framed as an abstract representation of a process or reality, designed to simplify complexity for analysis or prediction. Technology is described as the application of scientific and engineering practices to create and use artefacts, combining material tools with social knowledge. Computation encompasses formal symbol manipulation, execution of algorithms, and quantitative analysis to solve problems beyond the reach of direct observation or experimentation. In the RS corpora, these concepts are predominantly treated through their computational and engineering aspects. Algorithms are discussed in terms of optimisation and implementation rather than as sites where design choices embed particular values. Models are treated as self-contained evaluative units, abstracted from the contexts in which

they operate. Technology is framed as an inevitable driver of progress, with little attention to whose priorities it serves or how benefits are distributed. Computation is presented as an efficiency measure, divorced from the purposes and implications of what is being computed. This selective uptake narrows the definitional scope found in the IEEE glossary. The ethical, civic, and participatory dimensions embedded in the broader definitions are largely absent, which is analytically relevant to epistemic agency, as the mechanisms shaping what knowledge is produced, prioritised, and excluded are less frequently articulated in ways that invite public scrutiny. In the Russian corpus, the pattern is compounded by a lack of conceptual integration between these technical terms and the social or political contexts of RS deployment, underscoring the degree to which instrumentalist framings limit the interpretive space for contesting system goals.

5.5 Risk & Mitigation

The language of risk in RS research often signals a conceptual shift from ethics as moral judgement to ethics as a procedural artefact embedded in design and regulation. This section examines how RS discourse intersects with broader dynamics of techno-legal containment and risk governance. Drawing on Beck's concept of the 'risk society' (1992), RS development can be understood as operating within a framework where risk is managed rather than prevented. Within this framing, risk is embedded in regulatory and design systems as a technical uncertainty to be minimised, calculated, and absorbed. Beck's thesis foregrounds the shift from external to manufactured risks, which is relevant for algorithmic systems that both manage uncertainty and generate epistemic,

cultural, and political risks. However, RS research vocabulary tends not to emphasise this reflexivity. Risk is frequently presented as a technical concern, addressed through measures such as predictive calibration, anonymisation, encryption, or federated learning (Friedman, 1996; Jannach & Adomavicius, 2016). These measures protect system integrity and performance but give less attention to participatory control or civic transparency. This techno-legal framing supports a mitigation-oriented discourse, where the central question is not what kinds of harm RS systems might generate or perpetuate, but how those harms can be neutralised without altering the system's architecture. In this framing, mitigation functions as a linguistic proxy for control. Terms like risk, disaster, and harm are positioned as system-level disruptions to be pre-empted, rather than as social outcomes requiring political scrutiny. The increase in mitigation terminology is notable in Russian RS discourse after 2014, where system design often emphasises national coherence, threat prevention, and symbolic stability. Ethical responsibility in such contexts becomes procedural, shifting focus from whether a system should be built or used to how its effects can be contained. This critique does not imply that risk management is purely symbolic or devoid of control. Contemporary regulatory instruments such as the EU AI Act and standards like ISO/IEC 42001 explicitly require concrete actions, including risk assessment, system modification, usage restrictions, or withdrawal, thereby operationalising governance through enforceable procedures. Risk management frameworks also function as coordinating mechanisms in contexts where ethical values vary across polities and where rights claims may conflict, making consensus on substantive moral principles difficult to achieve. In such settings, proceduralisation provides a means of adjudicating between competing values

without privileging a single normative position. The tension identified here is therefore not between ethics and regulation as such, but between risk management as a necessary governance mechanism and the tendency, within RS research discourse, to treat procedural compliance as sufficient ethical resolution.

Risk management frameworks also operate through iterative learning mechanisms, where assessments of whether risks have been effectively controlled inform the revision of thresholds, categories, and mitigation requirements. In contemporary AI governance, including the EU AI Act, such feedback loops are intended to support ongoing identification of emergent risks, insufficiently treated harms, and the need for additional safeguards. In this sense, procedural risk management can function as an epistemic infrastructure that stabilises shared reference points for ethical debate across institutional and political contexts. However, the corpus evidence shows that RS research discourse rarely reflects this learning function explicitly, tending instead to frame risk management as a closed cycle of assessment and compliance rather than as a site of collective reasoning about acceptable risk and societal impact.

This procedural framing does not imply that ethical refusal or withdrawal is entirely absent in corporate practice. Some firms have reported pausing, modifying, or rejecting specific AI applications on ethical or governance grounds, particularly in response to regulatory scrutiny in the EU context (Smyth, 2025). However, available evidence suggests that such disclosures are uneven and increasingly formalised, with ethical considerations more often communicated through compliance-oriented reporting structures rather than through explicit accounts of refusal or contestation. Recent analyses of corporate AI disclosure

indicate that transparency is frequently selective and declining in jurisdictions with weaker regulatory pressure, particularly in the United States, reinforcing the shift from ethical deliberation to risk-managed disclosure. In RS discourse, fairness is often addressed after deployment, through metrics applied to functioning systems (Xu et al., 2021; Han et al., 2022; Roh et al., 2021). Responsibility becomes linked to compliance: whether protocols are followed, risks acknowledged, and outputs made explainable. This approach can understate the normative dimensions of RS design, including the values embedded, the assumptions reinforced, and the epistemic positions included or excluded. Within this cluster, the words risk, harm, safety, likelihood, trust, and ethical mark a conceptual shift in RS discourse from moral to procedural responsibility. Viewed through a moral perspective, these terms invite deliberation about whether harms should be prevented and whether systems should be built at all. Within procedural discourse, they are reduced to operational categories within system management. Risk is defined as a variable to be calculated, harm as a disruption to be measured, safety as a property to be maintained, and likelihood as a probability to be modelled. Trust is reframed as user acceptance or interface satisfaction, and ethical becomes a compliance descriptor rather than a question of principle. This movement from prevention to management aligns with the drift from ethics as a space for normative reflection to ethics as a policy artefact embedded in design and regulation. As a result, the scope for contesting how these terms are defined or acted upon narrows, and their civic dimensions are absorbed into the technical vocabulary of RS development.

Term trends

In the English-language corpus, risk maintains the highest relative prominence among this set, peaking in the most recent period and serving as a central organising concept in RS discourse. Harm, safety, and likelihood occur far less frequently but follow similar upward patterns, with likelihood reaching its maximum in the mid-period before slightly declining. Trust is by far the most common term in relative terms, though its trajectory shows a steady decline over time. Ethical begins with very low frequency but grows markedly after 2019, signalling the emergence of procedural vocabulary. Degree and closeness scores position trust and risk as the most connected terms in the network, functioning as hubs that link otherwise separate clusters of technical, governance, and performance language. This connectivity positions these terms at the centre of technical and regulatory vocabularies, where they are more frequently associated with system operation than with civic contestation. In the Russian-language corpus, 'безопасность' [safety] becomes the most prominent term by the most recent period, holding the highest degree and closeness scores in the network. 'Риск' [risk] also rises substantially, while 'вред' [harm] and 'вероятность' [likelihood] show more moderate growth. 'Доверие' [trust] and 'этический' [ethical] remain comparatively less frequent but both gain ground after 2014. As in the English corpus, centrality measures tie these terms closely to technical and regulatory clusters, embedding them in discourses of infrastructure stability, threat prevention, and controlled variability rather than in spaces of public reasoning or participatory oversight.

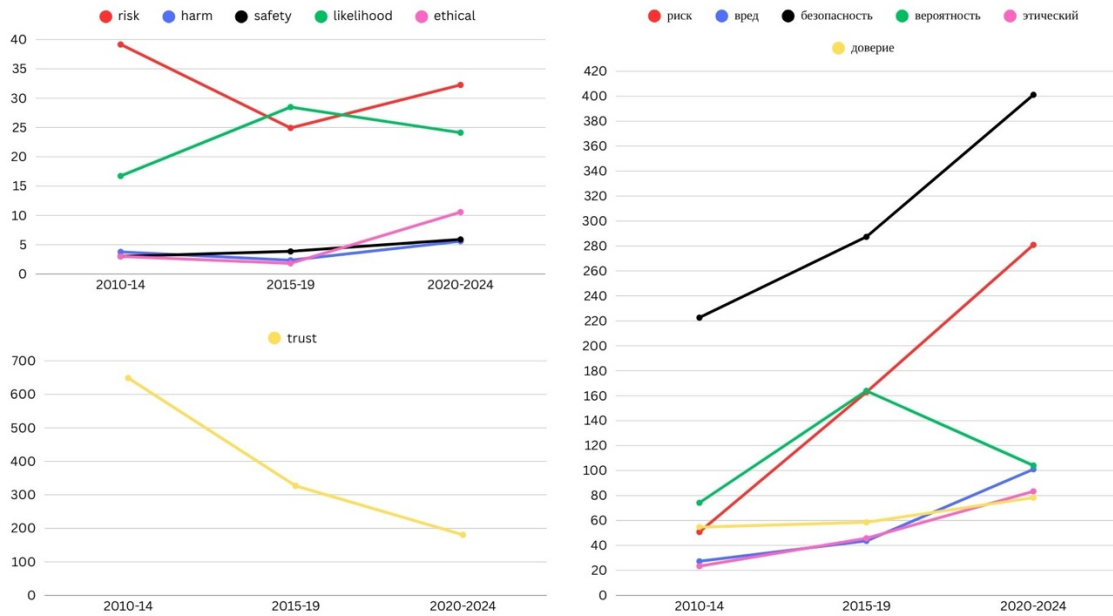


Figure 25 Normalised per-million growth of key Risk & Mitigation terms in the English corpus and in the Russian corpus.

Across both corpora, centrality patterns show that the most connected terms act as entry points into the technical vocabulary of RS systems, embedding risk, safety, and likelihood within operational logics. The language of harm shifts from civic contexts, where it could invite political scrutiny, into technical ones, where it is reframed as a system-level fault to be minimised. In this reframing, harm is defined through measurable disruption, risk as a quantifiable operational parameter, and safety as an infrastructural property rather than a civic condition. Trust remains lexically prominent in English but is framed through usability and acceptance, narrowing its civic scope. Ethical gains visibility but appears mainly as procedural, linked to compliance checks rather than open-ended reasoning. This reframing is analytically relevant to epistemic questions, as responsibility is more often framed procedurally through technical protocols rather than through relational or political terms. High centrality keeps these terms embedded in the technical core of RS discourse, limiting the scope for their reinterpretation in civic terms. Instead of debating whether a system should exist or whose knowledge it

validates, the discourse addresses whether risk is managed, harm is mitigated, safety is maintained, and trust is sustained at a usability level. In the Russian corpus, this narrowing is even more pronounced: ‘риск’ [risk] and ‘безопасность’ [safety] dominate technical framings, while ‘доверие’ [trust] and ‘этический’ [ethical] remain marginal, leaving little scope for civic or participatory dimensions.

Co-occurrence patterns show a concentration of these terms within technical and procedural frames. In the English corpus, risk most frequently co-occurs with user (69 to 192), privacy (42 to 169), and model (10 to 155), situating it between individual exposure, data protection, and system performance. Harm is linked to user (12 to 55), system (13 to 29), and model (0 to 26), reflecting a narrow operational framing. Safety appears with sustainability (0 to 36), system (0 to 18), and user (0 to 13), placing it largely in infrastructural and environmental contexts. Likelihood connects to user (36 to 291), model (27 to 174), and data (19 to 50), indicating its role in probabilistic modelling. Trust has the strongest co-occurrence profile, most often with user (1,350 to 2,224), system (382 to 658), and model (354 to 520), positioning it as a bridge between human-centred vocabulary and system design. Ethical is associated with user (1 to 56), system (1 to 56), and privacy (7 to 53), where it is tied to compliance and consent mechanisms. In the Russian corpus, ‘риск’ [risk] is most often linked to ‘система’ [system] (1 to 112), ‘безопасность’ [safety] (2 to 64), and ‘реализация’ [implementation] (0 to 23), embedding it in operational contexts. Вред [harm] connects to ‘моральный’ [moral] (0 to 88), ‘право’ [law] (4 to 24), and ‘здоровье’ [health] (2 to 23), signalling a more normative frame but without civic deliberation. ‘Безопасность’ [safety] co-occurs with ‘система’ [system] (6 to 215), ‘данные’

[data] (6 to 77), and ‘риск’ [risk] (2 to 64), consolidating its position as a technical safeguard. ‘Вероятность’ [likelihood] links to ‘пользователь’ [user] (4 to 49), ‘система’ [system] (2 to 25), and ‘данные’ [data] (2 to 20), reinforcing its modelling function.

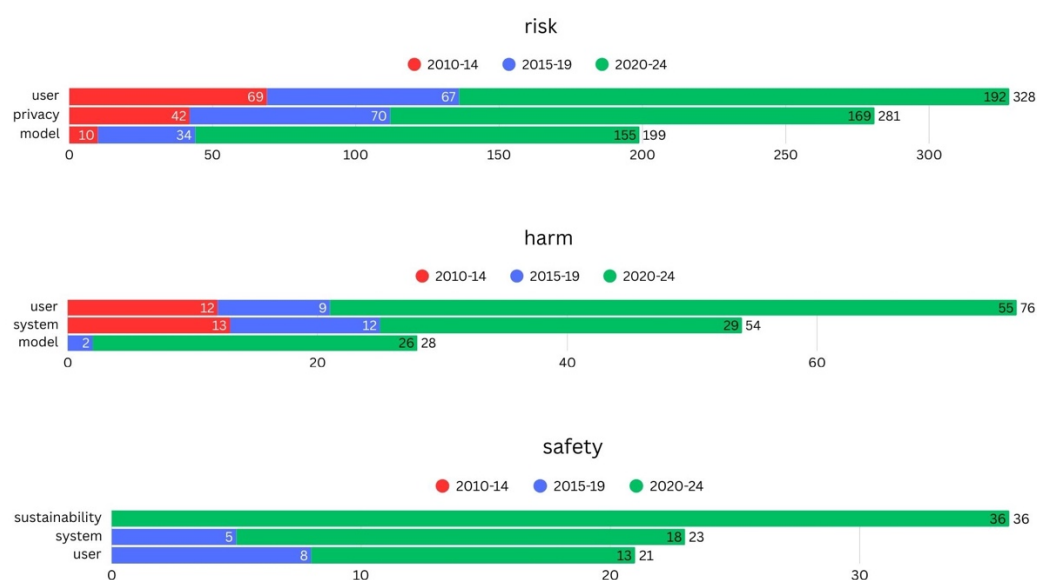


Figure 26 Top co-occurring terms for risk, harm, safety.



Figure 27 Top co-occurring terms for likelihood, trust, and ethical.

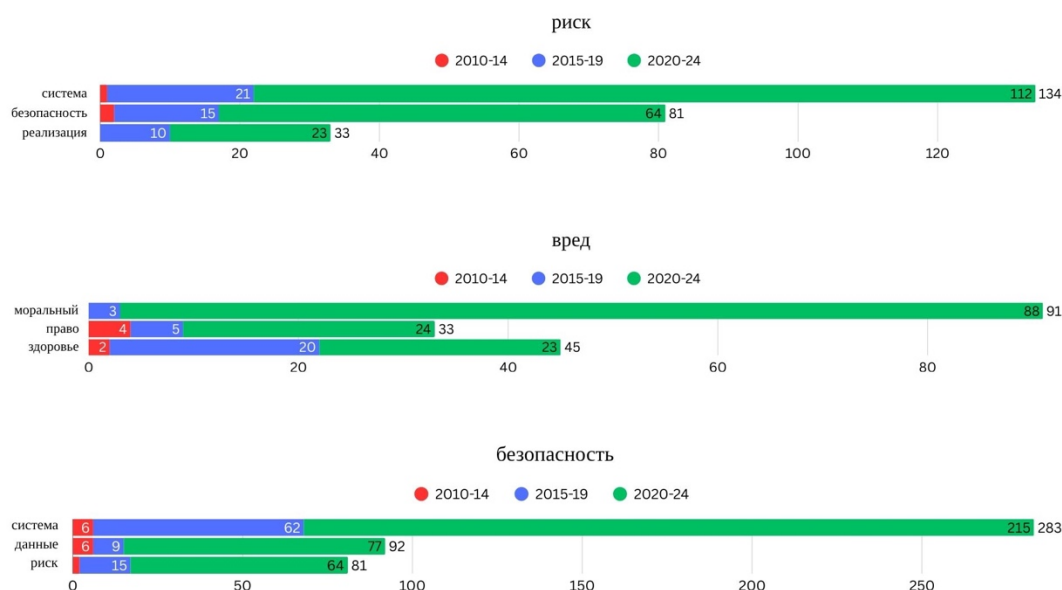


Figure 28 Top co-occurring terms for 'риск' [risk], 'вред' [harm], 'безопасность' [safety].

'Доверие' [trust] is associated with 'система' [system] (3 to 64), 'пользователь' [user] (0 to 10), and 'алгоритм' [algorithm] (0 to 10), situating it in controlled operational contexts. 'Этический' [ethical] connects to 'принцип' [principle] (4 to 39), 'система' [system] (1 to 32), and 'риск' [risk] (0 to 10), framing it as codified responsibility. Across both corpora, these co-occurrence patterns position risk, harm, safety, likelihood, trust, and ethical as connectors between human-facing and technical system language. Their placement within operational clusters corresponds to a discursive structure in which technical refinement is foregrounded and civic dimensions appear less frequently. These aggregate patterns are further illuminated in the following corpus examples, which show how each term's operational framing is enacted in specific RS research contexts across both languages.

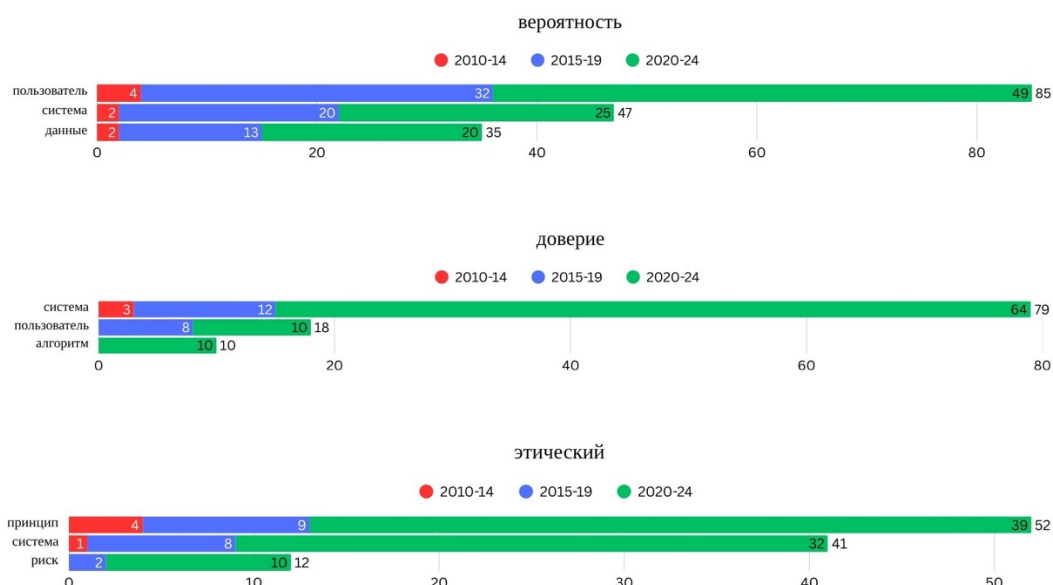


Figure 29 Top co-occurring terms for 'вероятность' [likelihood], 'доверие' [trust], and 'этический' [ethical].

Corpus examples

In the English corpus, risk and user co-occurrences illustrate a shift in emphasis across periods, from privacy-oriented containment toward engineering optimisation. In 2010–2014, a statement such as: 'User personal data, however, is not included in public datasets for research due to the risk of exposing user identities' frames risk as a procedural safeguard against identity exposure. By 2015–2019, 'Recommendation risk is defined as the user-to-list mismatch that leads to reduced satisfaction' presents risk as a performance metric. In 2020–2024, 'There is a risk of user privacy leakage; in future research work, the complexity of deep learning algorithms can be further reduced and fault tolerance improved' situates risk management within system refinement tasks. For harm and model, both periods point to an optimisation focus. In 2015–2019, 'Besides, the used categories are usually not designed for recommendation problem, which can seriously harm the recommendation model' treats harm as model degradation. In 2020–2024, 'Inaccurate consumption tags can harm model

optimization and performance' frames harm as an obstacle to be corrected within development cycles. Safety and sustainability appear in 2020–2024 as interlinked operational constraints: 'The top-k safe and mixed-k safe recommendations with safety as a hard constraint perform the best in customer safety with the lowest risk to customers and the lowest overexposure, and they also perform very well in business sustainability with very low exposure inequality.' Both terms are expressed through performance metrics rather than civic commitments. Likelihood and user connections remain closely tied to modelling objectives. In 2010–2014, 'Surprise is calculated as the difference between the likelihood that an item i is recommended for a user and the likelihood that an item i is suggested for any user' frames likelihood as a ranking calculation. In 2015–2019, 'For a given user set u , an item set j , and a set of user–item pairs that record previous interactions, the objective is to predict the likelihood of a user u interacting with an item i j ' abstracts the user into a predictive function. By 2020–2024, 'The predicted likelihood is calculated by taking the dot product of user and news representations' presents the concept as a computational operation. For ethical and system, 2020–2024 usage points to procedural embedding: 'Responsible recommender systems, on the other hand, concentrate on aligning system operations with ethical principles and social responsibilities, often focusing on fairness, accountability, and mitigating biases.' Here, ethics is codified as a system attribute rather than posed as an open question.

In the Russian corpus, 'риск' [risk] and 'система' [system] co-occurrences align with infrastructural framing. In 2015–2019, 'В настоящее время в России автономный автомобильный транспорт не зарекомендовал себя, поэтому

была выбрана рекомендательная система для исследования, которая позволяет уменьшить риск попадания в нештатную ситуацию' [At present, autonomous automotive transport in Russia has not proven itself, so a recommender system was chosen for study as it reduces the risk of entering an abnormal situation] situates risk within operational safety. In 2020–2024, 'Во избежание рисков следует осуществить конкретные организационно-управленческие решения, которые представлены в таблице. Таблица: способы уменьшения негативных последствий и предупреждения рисков... система защиты активно использовать эффективность страховой' [To avoid risks, specific organisational and managerial decisions should be implemented, as presented in the table. Table: ways to reduce negative consequences and prevent risks... the protection system should actively use insurance efficiency] embeds risk management in organisational safeguards. Other examples link risk to technical formalisation, as in 'Мы опирались на данные особой предметной области компьютерные науки, информационные технологии и информационная безопасность, которые достаточно органично поддаются формализации' [We relied on data from the specific domain of computer science, information technology, and information security, which lends itself quite organically to formalisation], and to transparency-driven optimisation, as in 'Чем прозрачнее рекомендательная система, тем более сильное положительное влияние она оказывает на пользовательский опыт, тем успешнее помогает в установлении контактов и повышает доверие учащихся' [The more transparent the recommender system, the more positive its impact on user experience, the more successfully it helps to establish connections and increases student trust]. For этический, usage in governance contexts reflects

standard-setting: 'Создание и продвижение общей дипломатической культуры является ключевым элементом формирования единой команды, включает в себя установление общих целей и ценностей, разработку этических стандартов и утверждение общих подходов к решению дипломатических задач' [Creating and promoting a common diplomatic culture is a key element in building a unified team, including the establishment of shared goals and values, the development of ethical standards, and the adoption of common approaches to addressing diplomatic tasks] frames ethics as a structured, procedural element of organisational cohesion. Across both corpora, these examples illustrate the central finding that key terms in the Risk & Mitigation cluster are embedded in operational and regulatory vocabularies. Their use privileges procedural compliance and technical optimisation, limiting their role as concepts that can invite civic deliberation or normative challenge.

Comparative insights with IEEE Glossary

The IEEE glossary presents risk, harm, safety, likelihood, trust, and ethics through definitions that combine technical precision with philosophical and policy-oriented breadth. In risk, the computational and engineering domains focus on measurable variables such as probability and consequence, while the ethical and philosophical strands retain the language of uncertainty, decision making, and unwanted events. In the RS corpora, prevailing usage more frequently reflects engineering and computational perspectives, framing risk as a calculable operational property rather than as a site for civic deliberation. This corresponds with the co-occurrence evidence in which risk is embedded in optimisation and containment logics rather than in open-ended discussions of

systemic harm. For harm, the IEEE materials recognise tangible and intangible dimensions, including physical injury, emotional impact, and social cohesion. RS discourse narrows this scope to performance degradation, reflecting the engineering definition's focus on the injurious consequence of a fault or failure. The civic and cultural harms acknowledged in the ethical and policy strands are rarely visible, reinforcing the observed drift from human-centred accountability to system-centred optimisation. IEEE definitions of safety extend across human, technical, and environmental contexts, combining probabilistic modelling with organisational duties. In RS discourse, safety is most often linked to optimisation constraints, as in the use of safe recommendations or safety as a hard constraint. This mirrors the engineering and computational framing but gives little space to the ethical and policy view of safety as a public condition. The IEEE framing of likelihood is grounded in statistical reasoning, focusing on distributional modelling and parameter estimation. RS usage follows this trajectory, treating likelihood as a predictive function or a vector calculation. Ethical considerations of likelihood, such as the uncertainty inherent in socially significant predictions, are absent. The focus remains on probabilistic efficiency rather than on interpretive openness. IEEE presents trust as both a relational and a procedural concept, combining philosophical, legal, and computational interpretations. In the RS corpora, trust retains high frequency in English but is largely framed as usability or acceptance, in line with procedural strands of the IEEE definitions. This procedural emphasis sidelines the relational aspects that support negotiated legitimacy. In the Russian corpus, trust is linked to infrastructural stability and transparency rather than to participatory governance. In ethics, IEEE offers a balance between professional codes, philosophical reflection, and public

accountability. RS discourse reduces ethics to procedural alignment, treating it as compliance with fairness or bias mitigation protocols. This corresponds with the engineering definition's focus on codified standards in professional practice and omits the moral reasoning and institutional oversight emphasised in the ethical and policy strands. Across these terms, the comparison reveals a consistent narrowing. IEEE definitions maintain a plurality of perspectives, combining technical, philosophical, and civic domains. RS discourse privileges computational and engineering framings, which limits the range of questions that can be posed. The focus shifts from whether a system's operation is justifiable or inclusive to whether its risks are managed, its harms mitigated, its safety optimised, its likelihoods predicted, its trust maintained procedurally, and its ethics codified. This movement from definitions that can sustain civic contestation to usages that prioritise technical closure is analytically relevant to the civic empowerment gap identified throughout this dissertation.

5.6 Conclusion

The analysis in this chapter has traced how RS discourse in both English- and Russian-language corpora constructs governance, agency, technical instrumentalism, and risk through vocabularies that privilege procedural alignment over substantive deliberation. Across thematic clusters, the findings point to a convergence between research and regulation that reproduces a shared logic of procedural oversight, favouring measurable outcomes over deeper engagement with the civic and cultural consequences of RS systems. The critique advanced here is not directed at measurement as such, which falls

squarely within the remit of computer science and engineering and is indispensable to RS development and evaluation. Rather, it concerns the discursive role that measurement plays in delimiting deliberation. In the analysed corpora, quantified indicators and performance metrics tend to function as endpoints of ethical and governance discussion, rather than as inputs into broader normative or civic reasoning. The issue is therefore not whether risks, harms, or fairness can be measured, but whether those measurements are treated as sufficient resolution, or as one contribution among others to contested judgments about acceptable system behaviour. In governance and accountability, procedural closure emerges through the dominance of operational metrics and compliance markers. English-language discourse often embeds governance in legal–technical configurations that emphasise explainability, interoperability, and auditability. Russian-language discourse situates governance within infrastructures of stability and security, aligning it with national policy objectives. Despite these institutional differences, both contexts limit governance vocabulary to system-internal parameters, with little discursive space for open-ended civic negotiation about RS purposes, values, or impacts. The agency and autonomy cluster extends this procedural narrowing. English discourse often frames user agency through personalisation features, interface affordances, or consent mechanisms that operate within predetermined system architectures. Russian discourse presents autonomy in terms of infrastructure resilience and national self-reliance, focusing on the integrity of the RS environment rather than on individual choice. In both cases, agency is rendered conditional and system-bound, reducing users to managed variables rather than recognising them as co-constructors of the recommendation environment.

Technical instrumentalism reinforces this alignment by embedding social concepts in engineering logics. Terms such as fairness, diversity, and inclusion are translated into measurable indicators, often as post hoc refinements to functioning systems. The broader ethical or political implications of these concepts are displaced by optimisation priorities, with performance gains serving as proxies for normative adequacy. This process constrains the range of possible interventions: technical fixes are privileged, while systemic redesign or participatory governance remain marginal. The risk and mitigation analysis shows how this logic operates across the lifecycle of RS development and evaluation. Risk, harm, safety, likelihood, trust, and ethical are treated less as prompts for moral reflection than as operational categories to be calculated, modelled, and contained. The co-occurrence and centrality patterns reveal their integration into technical and regulatory clusters, where they function as connectors between human-facing and system-facing language. In both corpora, the language of harm moves from the civic domain, where it could invite political scrutiny, into the technical domain, where it becomes a performance fault to be minimised. Trust retains prominence in English but is reframed as usability or acceptance, while in Russian it is tied to infrastructural stability. Ethics is reduced to compliance with procedural standards. This convergence between research and regulation is not accidental. The regulatory turn analysed here does not negate the democratic character of instruments such as the EU AI Act, which emerged through extended processes of political negotiation, public consultation, and institutional deliberation. Rather, the critique concerns how the ethical vocabulary codified through these democratic processes is subsequently absorbed into RS research discourse. Once translated into technical and

procedural language, ethical concepts tend to function as stabilising endpoints rather than as prompts for continued contestation. The issue, therefore, is not the legitimacy of regulatory ethics, but the way their discursive uptake can narrow the space for ongoing democratic interpretation and revision within research practice. It reflects shared pressures from commercial viability, regulatory harmonisation, and institutional risk management. By embedding civic concepts within system performance vocabularies, both research communities stabilise RS as an object of technical governance rather than as a site of democratic contestation. The result is a narrowing of epistemic agency: users, policymakers, and affected communities are positioned as recipients of protective measures, not as active participants in shaping the principles or purposes of RS systems. The IEEE glossary comparison underscores this shift. Where IEEE definitions sustain multiple perspectives, including technical, philosophical, and civic, RS discourse consistently privileges computational and engineering framings. In doing so, it collapses the conceptual breadth of key terms into procedural checklists. Risk becomes a quantifiable operational parameter, harm a disruption to be corrected, safety a constraint to be optimised, likelihood a probability function, trust a maintained usability level, and ethics a codified policy artefact. Each retains its ethical vocabulary but loses the contestability that could connect it to pluralist civic debate. The epistemic consequences of this shift are significant. When procedural closure defines the language of RS, the conditions for public reasoning about its purposes are diminished. The discursive space in which competing values, interpretations, and experiences could be articulated is replaced by a system logic that equates good governance with compliance, and ethical responsibility with adherence to technical standards. This creates a

linguistic environment where the empowerment gap is reinforced: those with access to the technical core define the meaning and scope of civic concepts, while others engage only with their operationalised forms. Chapter 5 displays how across both linguistic and institutional contexts the consolidation of RS discourse around procedural oversight constrains the possibility of substantive engagement with its societal role. By framing governance, agency, and ethics as performance conditions, and by embedding risk and harm in containment logics, RS research aligns itself with regulatory structures that favour stability over contestation. The convergence between research and regulation closes off opportunities for deliberative participation, narrows the range of possible interventions, and limits the capacity for RS systems to be shaped by diverse epistemic positions. This procedural consensus may secure system functionality, but it does so at the cost of diminishing the civic and democratic potential of recommendation environments.

Chapter 6: Epistemic Legacies in Algorithmic Research

In the previous chapters, RS were examined from three perspectives: Chapter 3 focused on user impact by analysing recommendations on video-hosting platforms YouTube and Dzen; Chapter 4 examined governmental AI policy frameworks in the EU, USA, and Russia; Chapter 5 analysed RS researcher discourse in English and Russian. In each, RS were explored as tools built to optimise platform-defined metrics such as clicks, watch-time, or engagement, used by developers to operationalise concepts of user choice and satisfaction (Nguyen et al., 2018). However, as these chapters and the preceding literature review illustrate, the first step in RS development is data reduction through abstraction and filtering of the training dataset. Techniques such as loss functions, anomaly smoothing, and ranking filters discard the significant long tail of data, due to it being deemed undesirable because it increases training time and costs, and, therefore, lowers the scores on common evaluation metrics. These steps train models to treat segments of information as having low relevance even before any recommendation is generated. Optimisation in this context requires prioritisation: the inclusion of one data point leads to the exclusion of another. The hierarchy is not neutral, as it is learned from historical distributions and operationalised through cost-sensitive thresholds defined by developers (Elkan, 2001). The result is algorithmic omission, where data considered irrelevant or costly to process is systematically excluded. What tends to be omitted are infrequent user behaviours, minority perspectives, or content with weak engagement signals. The result is an absence not just of information, but of alternative viewpoints and discourse, which impacts epistemic diversity and ability of a citizen to make a well-informed decision. Rather than serving as

a neutral optimisation routine, the system becomes a cultural technology that filters knowledge according to predefined efficiency metrics. The epistemic effect of omission can be understood through Foucault's account of archival power, where absence is not accidental but a mechanism that structures discourse (Foucault, 2013 [1972]). RS omit data in ways comparable to what Bowker and Star described as residual categories, those elements that classification systems are not built to accommodate (Bowker & Star, 1999). Exclusion operates systematically, narrowing the scope of knowledge long before users encounter outputs. Similar patterns appear in library catalogues, where rare languages or non-canonical texts were historically underrepresented (Olson, 2002), and in statistical sampling, where infrequent events were discarded to simplify analysis (Desrosières, 1998). The epistemic impact is not simply missing information but the reorganisation of what counts as relevant knowledge. This situates RS within a longer history of tools that privilege efficiency over range and render minority perspectives structurally invisible (Gitelman, 2013). This chapter does not introduce theory as an afterthought, but makes explicit the theoretical lenses that implicitly structured the empirical analyses in Chapters 3–5. It shifts from thematic organisation to analytic explanation by using theory to interpret how observed patterns persist, travel, and stabilise across technical, regulatory, and research contexts.

This chapter explores RS as cultural artefacts that inherit and reproduce historical patterns of epistemic control. Drawing on user, regulatory and research perspectives, it explores the technical routes by which statistical conventions move across domains, the principles that align optimisation with power, and the

economic and political constraints that solidify those alignments. By comparing English-language market-oriented RS artefacts with Russian state-directed analogues, the chapter highlights a central contrast: different institutional settings can converge on similar outcomes of reduced epistemic agency and managed plurality. This chapter maps the legacies and assumptions that precede any code, demonstrating that design choices reflect prior commitments rather than neutral logic. This perspective links data practices to philosophical questions about knowledge, authority, and public voice. RS are examined as a technology within a longer lineage of tools that filter, rank, and withhold information. Earlier classificatory infrastructures illustrate the continuity. Census systems reduced diverse populations to numerical aggregates that shaped taxation and representation. Bibliographic catalogues prioritised standardised entries that obscured local or marginal knowledge traditions. Broadcast schedules organised access to information through fixed time slots, privileging formats that fit institutional routines. Each of these media filtered knowledge to optimise for governance or market efficiency. RS extend this lineage by filtering at scale through statistical optimisation. The epistemic consequence is consistent across cases. Efficiency criteria displace pluralism and anchor communication in institutional priorities rather than in deliberative value.

RS inherit choices coded long before a single user has access to them. To navigate those choices, this chapter uses five analytic perspectives, each selected for its ability to expose a distinct layer of legacy that shapes what a contemporary RS is able to show or withhold. *Path-dependence* is drawn from science and technology studies because it explains how early design decisions

impact and transfer into default practice (Mahoney, 2000). *Instrumental rationality*, from political philosophy's critique of technical reason (Weber, 2019) highlights how optimisation metrics replace judgement with quantifiable goals, such as click-through or engagement-time. *Boundary-object translation*, from sociology of knowledge, maps how terms move across communities without shared meaning (Star & Griesemer, 1989). *Technological solutionism*, from critical studies of innovation cycles (Morozov, 2013), highlights how complex social or political problems are reframed as issues to be solved through technical fixes. Finally, *Economic lock-in*, from the history of technology, shows how revenue streams or command priorities define a design (David, 1985; Hughes, 1987). Discussed together, the five perspectives form a chain: early design choices (path-dependence) set the optimisation metrics (instrumental rationality), shared but vague terms (boundary objects) carry those metrics into policy, algorithmic fixes (solutionism) patch symptoms without touching the metrics, and revenue or command structures (lock-in) then solidify the whole cycle. These perspectives are not applied descriptively but as explanatory devices. Path dependence explains why specific optimisation routines recur across platforms; instrumental rationality explains why engagement metrics displace deliberative values; boundary-object translation explains why regulatory and technical interventions fail to realign objectives; technological solutionism explains why critique is absorbed as technical patches; and economic lock-in explains why these configurations persist despite sustained criticism. Together, they account for the gap between empirical observations and stronger causal claims, allowing the findings to be interpreted as tendencies and constraints rather than direct effects.

Application of these perspectives across English and Russian contexts displays parallel approaches to control. English-language RS discourse speaks of ‘models,’ ‘sustainability,’ and ‘mitigation,’ while compressing diversity into segments that maximise engagement in the interest of platform revenue and operational efficiency. Russian RS discourse deploys the language of ‘контроль’ [control], which limits content that diverges from the dominant narrative to preserve coherence. In both cases, users encounter a curated absence: not just fewer options, but a focus on what the system’s predefined parameters at individual or population level classify as understandable, profitable, or permissible. By tracing how statistical conventions, optimisation logics, semantic drifts, iterative fixes, and locked-in incentives align, the chapter positions RS within a historic line of information technologies such as catalogues, card files, and broadcast schedules, which organised knowledge to serve institutional objectives. This builds on Bowker and Star’s analysis of classification systems as infrastructures that ‘become relatively invisible without losing any of that power’ to shape what becomes visible, actionable, or excluded (Bowker & Star, 1999). Recognising these continuities shifts the debate from technical performance to epistemic design: RS do not just personalise, they execute inherited and affecting democratic process trends about which speech is disseminated, which signals matter, and how uncertainty should be resolved. To demonstrate the symmetry between English and Russian RS logics in contrasting political-economic environments, each analytic perspective is discussed separately. The empirical findings on linguistic diversity, regulation framework vocabulary, and research discourse from Chapters 3, 4, and 5 provide

the evidence for this exploration across intersecting and complementary perspectives.

Path-dependence argument

Path dependence explains why technical routines persist even when the context changes. The term was defined ‘as an idea that a small initial advantage or a few minor random shocks along the way could alter the course of history’ (David, 1985). In modern interpretations, ‘path dependence means that current and future states, actions, or decisions depend on the path of previous states, actions, or decisions’ (Page, 2006). James Mahoney described it as historical sequences in which accidental early events lock systems into self-reproducing trajectories (Mahoney, 2000). In social science, path dependence is used analytically to trace how feedback, contingency, and timing produce durable outcomes (Page, 2006; Garud & Karnoe, 2013; Liebowitz & Margolis, 1995; Pierson, 2000). In contrast, policy and governance literature tends to use path dependency descriptively to explain why institutional arrangements are difficult to reform, focusing on legacy systems, sunk costs, or political constraints (Kay, 2005, 2003; Greener, 2002; Augsdorfer, 2005; Pierson, 2000). Both terms refer to the same causal logic, but disciplinary priorities shape how the concept is applied: one to analyse formation, the other to interpret constraint. This chapter uses the social science framing. Path dependence provides an analytical framework for examining how RS design may inherit routines developed in other domains, where techniques can become embedded with limited contextual reassessment. It helps explain not only why such routines persist, but how they structure what RS can surface, suppress, or ignore.

A clear example of path dependence can be traced in the migration of Monte Carlo simulation from nuclear physics into financial modelling and, later, machine learning. The method was first formalised during the Manhattan Project in the 1940s, where it was developed to support neutron-diffusion calculations for nuclear chain reactions (Sood et al., 2021; Benov, 2016). The challenge was to approximate complex, multidimensional integrals that lacked analytical solutions. To solve this, scientists used random numbers to simulate particle interactions and estimate system outcomes probabilistically. The term 'Monte Carlo' was introduced by physicists at Los Alamos, reportedly referencing the Monte Carlo Casino due to the role of chance in the process (Metropolis & Ulam, 1949; von Neumann & Ulam, 2024). The technique drew on earlier work by von Neumann and Ulam, who had explored stochastic methods for systems resistant to deterministic modelling. Its core assumption is that many independent random trials will converge on a stable expectation, a principle aligned with the statistical mechanics of chain reactions and equilibrium behaviour. This made it well suited to nuclear applications, where large numbers of atomic events follow predictable aggregate patterns. Three decades later, the same routine entered financial modelling. Boyle used it to estimate option values by simulating future prices, assuming they would follow a log-normal distribution (Boyle, 1977). The approach applied the same statistical logic to asset price dynamics, based on the assumption that markets behave as stochastic systems with stable probabilistic rules. The algorithm remained largely unchanged across domains, though specific adaptations emerged, such as volatility adjustments and discounting techniques. The transition was enabled by the belief that random

sampling could yield accurate approximations in any domain governed by probability. However, this assumption overlooked key contextual differences. While nuclear processes conform to physical laws and known probabilities, financial markets are shaped by regulation, feedback loops, and human behaviour.

Nevertheless, the technique's continued use contributed to pricing models that carried an appearance of quantitative authority. The leap from physics to finance was not the end of this contagion effect. Monte Carlo methods were adopted across domains that require estimation under uncertainty. For example, AlphaGo applied a similar technique to simulate future moves and outcomes, using these projections to inform its decision-making based on estimated probabilities of success (Fu, 2016). Monte Carlo Tree Search, used in AlphaGo and its successors, is a structured extension of Monte Carlo. It combines stochastic simulation with decision-tree traversal, using sample rollouts to estimate the value of actions. The logic remains the same: with repeated simulations, the average result approximates expected value. Today, Monte Carlo is used in search strategies for large-language-model agents (Świechowski et al., 2023; Preil & Krapp, 2022; Cowling, Powley & Whitehouse, 2012; Chaslot et al., 2008; Browne et al., 2012). In these cases, tree-of-thought and roll-out sampling generate token trajectories, then rank them using a value function based on prompt probability or reward. These outputs prioritise continuations that align with system-defined metrics. Statistical independence and convergence are assumed, although autoregressive generation causes drift over time and reinforcement learning modifies the search distribution during training. The

assumption from the original Monte Carlo experiments remains in place, while implementation is supported by software libraries and systems that mask contextual variation. Monte Carlo methods are used in computer graphics for light transport simulation, in epidemiology for modelling disease transmission, and in power systems for grid performance analysis. These applications share a reliance on probabilistic sampling, but the reliability of results depends on how the method is implemented and how well the system matches its assumptions. As shown in applications of Monte Carlo simulations reviewed by Chan, outcomes in reactor shielding calculations are sensitive to boundary assumptions (Chan, 2013). In climate forecasting, rare events are often excluded due to limits on sampling depth. In financial models, simplified volatility inputs shape risk projections. Across these cases, the Monte Carlo method functions as a general routine, while the outputs reflect both the sampling process and the framing decisions made in advance.

The transfer of probabilistic routines across domains illustrates how assumptions built for one environment become naturalised in another without re-evaluation. In physics, convergence rests on physical laws that govern particle interactions. In finance or social media, no equivalent laws exist, yet the same statistical logic is applied. The method can be read as carrying epistemic authority through its mathematical form rather than through contextual validation. Shannon's information theory travelled a similar path when imported from telecommunications into linguistics (Kolmogorov, 1956; Csiszár, 2002). Entropy, originally a measure of signal transmission, was adopted as a proxy for linguistic complexity, even though speech is embedded in context and culture. These

migrations illustrate how technical routines gain legitimacy by appearing universal while their assumptions are domain specific.

This continuity illustrates what Hughes called system ‘momentum,’ where once a technique integrates into training pipelines, ‘manufacturing corporations, public and private utilities, industrial and government research laboratories, investment and banking houses, sections of technical and scientific societies, departments in educational institutions, and regulatory bodies add greatly to the momentum’ (Hughes, 1987). Momentum arises not only from technical performance but from alignment across institutions, standards, and expectations. Once a method is installed, it becomes embedded in professional training, procurement policies, and software dependencies. Historical examples include the persistence of the QWERTY keyboard layout and the continued use of floating-point arithmetic standards, even when alternative numeric formats reduce rounding error or improve interpretability (Brain et al., 2015). Certification frameworks, licensing protocols, and Application Programming Interface (API) design further entrench specific implementations (Bloch, 2006). Substitution, even when viable, often requires rewriting system interfaces, retraining personnel, and renegotiating compliance, making change structurally costly. As a result, methods remain in place long after the original rationale fades, maintained by the wider system rather than continual re-assessment.

Comparable persistence is evident in railway gauges that determined cross-border interoperability, in TCP/IP protocols that stabilised global networking, and in Unicode standards that fixed representational codes for digital

text. In each case, once adoption reached a threshold, alternatives became impractical despite technical merits. RS metrics such as click-through rate operate in the same way. They gain institutional weight through procurement policies, software libraries, and contractual agreements. Change requires renegotiating the entire surrounding infrastructure, making persistence the rational option even when epistemic diversity is compromised.

RS carry this momentum. Collaborative filtering uses matrix factorisation methods originally developed for solving dense linear systems in physics and engineering. In RS, they are applied to user–item interaction matrices to uncover underlying dimensions that predict unseen preferences. The original methods assume complete data and stable correlations. In RS, the data is sparse, non-randomly missing, and primarily shaped by human behaviour, though constrained by system exposure. Factorisation remains dominant because it is compatible with efficient numerical computation, matches standard performance metrics such as root-mean-square error, and is available in commonly used software libraries. Bayesian bandits are based on decision strategies from probability theory and gambling models such as the multi-armed bandit problem (Scott, 2010). In RS, they are used to select which items to show to users under uncertainty. The original assumption is that reward distributions are independent and stationary. RS environments challenge this assumption: user preferences shift, item exposure changes behaviour, and feedback loops distort observed outcomes. Despite this, bandit algorithms remain standard due to their compatibility with online experimentation and constrained optimisation setups. K-nearest neighbour methods come from early classification tasks in pattern

recognition, where the idea is that items close in feature space are likely to share labels (Fraiman, Justel & Svarc, 2010; Ahmed et al., 2021). In RS, these methods are used to recommend items similar to those a user has interacted with, or to identify users with similar histories. The original assumption is that similarity is stable and meaningful across the input space. In RS, where preferences are context-dependent and histories are partial, this assumption introduces bias towards users with more interactions or dominant categories. Graph-based methods originate in network analysis, where centrality and connectivity are used to identify influence or importance (Mesbahi & Egerstedt, 2010). In RS, they are used in trust-aware recommenders or item relationship models (Guo et al., 2020; Gao et al., 2023; Deng, 2022). These approaches assume the graph is stable, links are meaningful, and data reflects structural proximity. In practice, RS graphs are constructed from interaction logs that reflect system design choices rather than underlying relationships, introducing structural artefacts. Embedding models come from natural language processing, where words are mapped to vectors based on their co-occurrence in fixed windows (Li & Yang, 2018). In RS, the same techniques are used to learn item or user embeddings based on interaction sequences. The original assumption is that proximity in sequence implies similarity. In RS, this collapses diverse user actions into a single metric, ignoring intent, context, or platform-specific logic. These techniques remain in place because they are implemented in standard software libraries, supported by hardware-accelerated pipelines, and validated against metrics that reward short-term predictive accuracy. This creates a path-dependent system in which method choice determines input processing, optimisation targets, and the type of feedback captured. The result is an RS architecture that is more strongly

oriented around technical routines imported from other domains than around explicitly articulated user needs. Each imported method is commonly used in RS development and brings its own built-in priorities, favouring dense signals over sparse ones and short-term rewards over long-term uncertainty. This creates a path-dependent architecture that favours historical popularity and discourages novelty, shaping outcomes even before personalisation begins.

Technical-instrumental rationality

Technical-instrumental rationality can be used to analyse situations in which techniques such as algorithms, optimisation routines, or evaluation metrics are treated as self-justifying ends, even when original goals are stated in formal definitions. The term originates in Weber's theory of rationalisation, where he distinguished between *Zweckrationalität* (instrumental rationality) and *Wertrationalität* (value rationality) (Weber, 2019). Instrumental rationality refers to actions oriented toward the most efficient realisation of a given goal through rule-based calculation. In contrast, value rationality is guided by beliefs about intrinsic worth, independent of outcomes. Weber associated instrumental rationality with bureaucratic and capitalist modernity, noting that the growth of formal procedures often eclipses normative reasoning (Weber, 1978). This distinction was later developed by Horkheimer and Adorno, who argued that instrumental reason, when used without reflection, enables domination under the guise of technical progress (Horkheimer, Adorno & Noeri, 2002). For instance, workplace surveillance tools are introduced as efficiency measures but end up reinforcing managerial control over workers. Habermas extended this by showing how systems replace public debate with technical procedures, turning

communicative goals into calculative ones (Habermas, 1985). For instance, algorithmic content moderation turns questions of harm and fairness into threshold tuning, bypassing deliberation about values or social impact. In digital systems, technical-instrumental rationality describes the practice of calculating and optimising the most efficient means through formal rules, algorithms, and technical control, without evaluating what ends those means are designed to serve, such as shaping attention, structuring discourse, or guiding decision-making (Mueller, 1979; Whimster & Lash, 2014). Using this principal, Winner notes that large systems drift toward maximising measurable outputs, reducing qualitative goals to numerical proxies (Winner, 1986).

In Chapter 5, this proxy function is illustrated through the language used to describe the optimisation metrics. Once installed, the metric drives design and interface choices, while questions about civic participation and agency fall outside the optimisation loop. The logic is circular. A platform specifies a reward *X*. Engineers search parameter space to maximise *X* using, for instance, loss function. The observed increase in *X* justifies the metric's authority. Content that raises engagement but potentially narrows discourse is classified as 'relevant,' while material that broadens perspective but depresses clicks is defined as noise. Efficiency stands in for meaning, even when the outcome undermines democratic processes. Instead of evaluating what forms of knowledge a system should support, platforms define success through metrics such as watch-time or click-through. These metrics function as proxies for relevance, prioritising content that generates measurable interaction. Content expressing complex or minority perspectives typically underperforms on such metrics due to lower immediate

engagement (Akpinar & Fazelpour, 2024). As a result, optimisation processes favour content with high interaction rates, regardless of its epistemic diversity. This produces an environment where information visibility is determined by performance scores rather than public significance.

Winner argues that what such systems do is reorganise power without explicit consent:

Because choices tend to become strongly fixed in material equipment, economic investment, and social habit, the original flexibility vanishes for all practical purposes once the initial commitments are made. In that sense technological innovations are similar to legislative acts or political foundations that establish a framework for public order that will endure over many generations. For that reason the same careful attention one would give to the rules, roles, and relationships of politics must also be given to such things as the building of highways, the creation of television networks, and the tailoring of seemingly insignificant features on new machines. (Winner, 2020 [1986])

The parallel with bureaucratic quantification is clear. Nineteenth-century census procedures reduced identities to occupational or demographic categories that enabled taxation and military planning. Taxation systems translated property and labour into numerical forms that allowed comparison but excluded qualitative dimensions of life. RS replicate this logic by translating civic discourse into measurable signals such as clicks or session length. The authority of numbers stabilises decisions, even when the measures fail to capture substantive meaning. Instrumental rationality transforms heterogeneous practices into uniform scores that guide institutional decisions without public deliberation. Once technical decisions are embedded in infrastructure, they become difficult to reverse and operate without further scrutiny. The quote shows how systems stabilise early choices through material investment and routine use. For instance,

a RS model trained on engagement remains aligned with that objective, even if platform goals change. This stability is not neutral. It governs how visibility is distributed, which content types are amplified, and what user actions are reinforced. The comparison to legislative acts in the quote illustrates that systems shape public order by structuring interaction, not by regulating speech directly. These effects persist over time because technical systems rarely include mechanisms for collective review or adjustment. From this perspective, political impact can be understood as operating through design choices rather than through deliberative processes.

RS exemplify this shift. Ranking weights encoded in a matrix determine which voices surface and which disappear, but the weighting process is opaque to users and regulators. The technical frame casts the decision as neutral: a gradient descent step, a learning-rate schedule. In practice it reallocates attention, amplifying speakers who match historical engagement patterns and silencing those who do not. The power shift is structural, not conspiratorial, and it arises because optimisation abstracts away the social purpose of communication. Winner also states that 'of all the computer enthusiasts' political ideas, there is none more poignant than the faith that the computer is destined to become a potent equalizer in modern society,' a claim which is still applicable in the discourse of the large language model agents (p.112). Technical-instrumental rationality also explains cross-context similarity. Western platforms optimise revenue by aligning X metric with advertising strategies, as shown in Chapter 3 in the context of demographic and gender prioritisation. Meanwhile, Russian platforms optimise stability by aligning X with state guidelines. Both

pursue efficiency relative to their objective, and both treat divergence from that objective as system error. Technical teams optimise loss curves, overlooking broader consequences, as the metric provides an explicit reference point that organises internal coordination. In this context, the metric's authority pre-empts debate. Users have agency to choose what to watch, but not which objective the system should serve. Regulators audit outputs or documentation, not the premise that engagement ought to rule. Winner called this 'technological somnambulism,' where society walks into new power relations because optimisation appears apolitical (p.5).

RS display this pattern at scale. Technical-instrumental rationality functions as a design constraint that replaces normative judgement with quantifiable optimisation. Designers select a reward signal such as watch time, click through rate or dwell time based on measurability rather than epistemic value. System success is defined by metric improvement, closing the loop between design input and behavioural output. This pattern is not limited to commercial systems. It characterises supervised machine learning pipelines used in content selection. A model is trained to minimise loss on objectives drawn from behavioural history. Predicted relevance becomes a proxy for expected interaction and not for public significance. The model reinforces historical interaction patterns and stabilises them through optimisation. YouTube's RS uses deep models trained on behavioural features such as session duration, rewatch rate and skip behaviour (Davidson et al., 2010). These features are aligned with engagement targets. The system promotes content that correlates with repeat interaction regardless of civic value. Content that produces consistent

responses gains visibility even if it fragments discourse or simplifies reasoning. System scale prevents consistent human oversight. Retraining depends on aggregate metrics such as total watch time or daily active users. Metric improvement is treated as success. A communication system becomes an agent of behavioural regulation. This structure shapes information access. Content with low immediate engagement such as policy analysis, minority perspectives or critical journalism is deprioritised. It underperforms against the metric and becomes statistically marginal. Visibility depends on alignment with past interaction data rather than on relevance to civic participation. Optimisation coherence reinforces exclusion. Engineering, product and policy teams coordinate around fixed metrics. Alternative objectives such as diversity or epistemic range are excluded if they cannot be integrated into the optimisation routine. Once infrastructure stabilises, changing success definitions becomes institutionally difficult. Users do not shape ranking logic. Regulators review outputs but do not access the optimisation core. Efficiency benchmarks preserve engagement as the system's central aim. Personalised optimisation segments audiences. RS generate distinct content hierarchies per user. Civic discourse, which depends on shared exposure, can be interpreted as becoming increasingly fragmented under personalised optimisation regimes. The metric defines system evolution. A model that increases watch time is classified as improved even if it deprioritises deliberative content. RS remain stable under critique. Most interventions focus on moderation, explainability or transparency and do not challenge the optimisation loop. Engagement continues to serve as the reference for success. RS operate as infrastructures of epistemic distribution shaped by technical-instrumental rationality. Their political effects result from cumulative

optimisation outcomes. Filtering, ranking and visibility are governed by performance metrics and not by civic discourse about public relevance.

Boundary-object translation

Boundary-object translation, introduced by Star and Griesemer, describes how artefacts move across social domains while adapting to local contexts (Star & Griesemer, 1989). Boundary objects retain enough flexibility to satisfy different demands without requiring consensus. Their advantage is coordination, while their risk is semantic deviation. The concept originated in the sociology of science to explain how collaboration occurs between actors with incompatible assumptions. Star and Griesemer analysed the Museum of Vertebrate Zoology, where scientists, administrators, and amateur collectors worked together without a shared epistemic framework. Each group applied different standards of classification, documentation, and purpose. Scientists prioritised taxonomic rigour, administrators managed funding and institutional continuity, and collectors contributed local specimens using informal methods. They did not agree on what constituted valid knowledge but still cooperated using shared artefacts such as catalogues, forms, and specimen labels. These artefacts served as points of coordination, allowing each group to interpret them differently while continuing joint work. The boundary objects functioned not by creating agreement but by supporting parallel interpretations. This allowed structured collaboration without conceptual alignment. This section does not describe individual engineers or policymakers as homogeneous groups. It analyses institutional logics, evaluation regimes, and organisational incentives that structure how technical and regulatory actors interact in RS governance. The

contrast drawn here concerns artefacts, accountability mechanisms, and success criteria within professional fields, not the intentions, expertise, or lived experience of practitioners, which vary widely across roles and organisations.

RS governance and discourse, discussed in Chapters 4 and 5, illustrate both constraints of boundary-object translation: coordination and semantic drift. Regulatory and engineering communities use shared terms to facilitate dialogue, but the meanings of those terms diverge significantly across contexts. ‘Model’ and ‘risk’ are central examples. In regulatory texts, a model refers to a legally significant artefact whose outputs must comply with rights frameworks. It is a target of compliance scrutiny and may trigger accountability proceedings. In engineering, a model is a statistical object designed to generalise from data and optimise performance. In IEEE Ethically Aligned Design, it is defined as a simplified or abstract representation of reality (IEEE, 2017). In practice, it is a parameterised function trained to minimise a loss function and evaluated using task-specific metrics, with generalisation assessed via performance on unseen data. These meanings reflect distinct evaluative systems. For engineers, the model is judged by predictive validity and computational efficiency. For regulators, it is judged by normative consequences and legal compliance. The same term links two incompatible accountability structures. The contrast becomes clearer when the term ‘risk’ is introduced. In engineering, risk is formalised as a measurable property of model behaviour. It is expressed through statistical proxies such as error rates or classification thresholds. The ROC (Receiver Operating Characteristic) curve, originally developed during World War II to measure radar signal detection accuracy (Yang & Berdine, 2017;

Murphy et al., 1987), is now used to evaluate how well a model distinguishes between classes across decision thresholds (Hoo, Candlish & Teare, 2017). Risk becomes a matter of calibration. If the curve is satisfactory, the model is considered to manage risk. In regulatory contexts, risk refers to potential harm to individuals and society. It includes discrimination, exclusion, manipulation, or other violations of fundamental rights. These cannot be reduced to statistical performance. The EU AI Act, for instance, defines unacceptable risk in terms of legal breaches and societal disruption. From this perspective, a model that performs well on ROC metrics may still be considered harmful. The divergence mirrors other domains. In aviation engineering, risk is defined through probabilistic models of component failure rates and mean-time-to-failure benchmarks (Woch, Zieja & Tomaszewska, 2017; Mili & Sheldon, 2009). In insurance law, risk means liability and financial exposure (Priest, 1986; Ewold, 1991). The same word is used across sectors, but the ways it is evaluated are very different. Because the term travels easily, it helps coordination while hiding these differences. RS governance works the same way. Technical and legal actors both talk about risk, but they measure and address it in different ways. Developers may assume that risk is addressed once technical indicators are met. Policymakers may view the same system as creating unacceptable outcomes. This conflict stems from different assumptions about what risk is and how it should be evaluated. The divergence is reinforced by the ambiguity of the term model in common usage. A model can be a toy, a blueprint, or a person. As it circulates through technical, legal, and public discourses, the term becomes increasingly unstable. This instability supports the boundary-object function. Each actor can use the term while applying their own assumptions. Coordination

is maintained through shared vocabulary, but there is no alignment on meaning. The object enables cooperation while concealing dissonance. The same pattern holds for the term risk. For engineers, it is an input to system design. For regulators, it is a public concern. Developers may introduce traceability tools or differential privacy techniques to show accountability. Legal frameworks define accountability as the duty to explain decisions and face consequences when norms are violated (Cornock, 2011). The IEEE glossaries reduce this to responsibility, omitting the requirement for sanctions or justification (IEEE, 2017). As a result, compliance becomes formal. Logging or audits satisfy technical criteria while systemic harms remain unexamined. Boundary-object translation explains the persistence of these gaps. The issue is not failed communication but structural misalignment, where boundary languages can miscommunicate by design. Terms such as model and risk operate across domains but carry incompatible meanings. Regulators expect transparency and normative justification. Engineers deliver statistical summaries and documentation. The boundary object sustains cooperation but prevents resolution. This structure remains in place as RS governance evolves. Without shared interpretive frameworks, the same vocabulary continues to obscure foundational disagreements. In RS development, this allows compliance procedures to reinforce existing optimisation goals rather than address their effects. This divergence has been extensively analysed in science and technology studies as a problem of coordination across fields with different evaluative logics, rather than as a mismatch between individual actors. Boundary objects enable cooperation without shared epistemic commitments, producing procedural alignment without conceptual agreement (Star & Griesemer, 1989).

The shared vocabulary, ironically, masks incompatible evaluative frames. Terms such as ‘model,’ ‘risk,’ and ‘accountability’ appear to align policy and technical discourse, but they operate within structurally different epistemologies. Policymakers use these terms to define normative goals. Their intent is to shape technological systems in ways that uphold fairness, transparency, and accountability across social contexts. This presumes that values can be stabilised through language and codified into design constraints. It also assumes that actors downstream will interpret these terms through their normative content. For engineers, however, these terms become variables, constraints, or metrics to be optimised. The underlying logic treats all frames as functionally equivalent: what matters is not what ‘accountability’ means in a legal or philosophical sense, but whether a system satisfies the conditions specified in the audit template. This divergence produces a boundary object that malfunctions. Each side believes it refers to the same object, but applies different criteria to evaluate whether the object meets its purpose. From the policymaker’s side, the object should reflect contextually grounded reasoning about social impact. From the engineer’s side, the object is part of a task to be completed within a bounded problem space. The policymaker’s framework assumes that harm is mediated by culture, context, and historical inequality. The engineer’s framework assumes that harm can be isolated, defined, and resolved through adjustments to a model’s parameters or architecture. This is not a matter of disagreement over detail, but of different ontologies. Policymakers ask whether the system’s logic supports pluralistic representation. Engineers ask whether the gradient update leads to a lower loss. Both use the word ‘accountability,’ but the structure of evaluation is not the same.

The result is a technical infrastructure that satisfies the letter of the law while bypassing its intent. Policymakers attempt to guarantee fairness by defining templates that formalise desirable outcomes: data audit trails, bias assessments, transparency reports. Engineers treat these templates as design briefs. If a fairness checklist includes logging procedures and differential privacy noise, those elements are added. The question of whether these additions meaningfully shift power or mitigate epistemic exclusion is treated as external to the design task. Fairness becomes something that can be passed like a unit test. Once the form is filled, the system is considered compliant. In this logic, context is removed. A fairness framework that was meant to reflect structural inequalities is converted into a set of technical instructions. If the algorithm passes, then the fairness has been ‘solved.’ In this interaction, the boundary object becomes false. The engineer recognises that policymakers intend the object to carry ethical weight, but deploys it strategically to satisfy surface-level requirements. This dynamic combines strategic ambiguity with performative compliance. It is not deception in the ordinary sense. It is a rational strategy within the incentive structures of each field. Engineers respond to what is measurable. If the accountability requirement is phrased as a demand for explainability, then a system log or decision tree approximation will suffice. Whether this satisfies broader normative goals is not part of the technical evaluation. From the policymaker’s side, this may appear as a malfunction. The boundary object was meant to bridge perspectives, but instead becomes an instrument of minimal compliance. This process mirrors Bourdieu’s account of field-specific capital (Bourdieu, 1983). Each domain brings its own logic, structure, and hierarchy. In the regulatory field, symbolic capital is drawn from legal coherence, democratic

legitimacy, and the protection of rights. In the engineering field, capital is accumulated through efficiency, scalability, and technical optimisation. When policy language enters the technical field, it is recontextualised. The framing remains, but the logic is altered. Policymakers see fairness as deliberative and situational. Engineers see it as a constraint to be met. As a result, fairness is implemented in form but not in substance. Templates are completed, audits passed, and metrics optimised. The incentive is to solve for the indicator, not to question its premises. This asymmetry is sustained by unequal visibility. Engineers often understand what policymakers want, at least at the level of vocabulary. Policymakers rarely see how the system interprets their input. This is a case of asymmetric recognition. The technical side understands the symbolic value of the regulatory terms and uses them to demonstrate alignment. The policy side assumes that formal alignment implies conceptual agreement. The engineer may comply with the checklist while maintaining the same optimisation routine. The result is a selective appearance of agreement that hides structural differences. The failure is not in communication, but in translation. The shared vocabulary gives the impression of mutual understanding. In practice, the terms operate differently across fields. Policymakers aim to shape behaviour through normative framing. Engineers treat the framing as a design constraint. Any term, once formalised into a template, produces the same outcome. For policy, the frame is context dependent. For engineering, the frame is input dependent. This difference ensures that RS systems remain formally compliant while epistemically unchanged. Meanwhile, RS discourse increasingly cites accountability and mitigation, terms rooted in administrative law such as data protection, anti-discrimination frameworks, and procedural standards for public

institutions. Researchers translate accountability into proxy metrics such as log retention or differential-privacy proofs, assuming auditability satisfies public oversight. Yet legal traditions define accountability not merely as responsibility, but as the obligation to explain and to face sanction if norms are breached (Cornock, 2011). The IEEE EAD glossary, however, diminishes accountability to responsibility alone (IEEE, 2017).

A system can log every gradient update and still evade liability if its objective function violates anti-discrimination norms. Traceability alone does not ensure normative compliance. This gap is discussed in Chapters 4 and 5, where regulatory frameworks are shown to prioritise corrective action, while research discourse treats technical auditability as sufficient. The field of explainable AI (XAI) illustrates this divergence. XAI methods such as saliency maps or SHAP (SHapley Additive exPlanations) values are designed to make model outputs more interpretable (Younisse, Ahmad & Abu Al-Haija, 2022; Roshan & Zafar, 2021; Al-Najjar et al., 2023). These tools provide local approximations of decision boundaries or feature importance scores, offering what appears to be transparency. However, they do not interrogate the objective function itself. They do not expose whether the model's core logic aligns with anti-discrimination norms. For example, a system trained to maximise engagement may consistently prioritise sensationalist content. XAI tools can show which input features contributed to a given prediction, but they do not reveal whether the optimisation goal itself is normatively acceptable. In practice, this results in explanations that clarify individual decisions without questioning the fairness of the system's overall behaviour. The boundary-object concept clarifies why harmonisation

stalls. Each community projects its own commitments onto portable terms and assumes mutual understanding. Regulatory texts assume that models reveal trade-offs that can be surfaced and contested. Developers assume that legal actors will accept quantitative proxies such as logs, feature attributions, or ROC curves. This mismatch becomes institutionalised when XAI tools are used as stand-ins for accountability. A developer may present SHAP plots to demonstrate transparency even when the model continues to produce discriminatory patterns. These patterns may be driven by biased training data or by objectives that prioritise engagement or retention. The outcome, observed in EU, USA, and Russian AI governance frameworks, is selective compliance. Engineers meet formal documentation requirements by adopting explainability tools. Substantive concerns such as echo chambers, epistemic exclusion, or the reinforcement of social hierarchies remain embedded in system design. Explanations are generated, but the optimisation logic stays unchanged. The boundary object supports formal cooperation while blocking structural challenge.

Terms such as ‘model’ and ‘risk’ function less as stable references and more as fragments of an exchange language. They circulate within a system where neither side accepts the responsibility implied by the other, nor cedes the authority that such acceptance would require. These terms enable procedural alignment without conceptual agreement. As Peter Galison notes, ‘what thrives in the interstitial zones is neither trivial nor purely instrumental, it is a form of scientific exchange language,’ shaped by local demands rather than shared understanding (Galison, 2010; Krige & Pestre, 2013). This shared but unstable language permits coordination, but it also enables forms of misrecognition. What

appears as consensus may in fact conceal structural asymmetries, where one side shapes the terms of engagement while the other interprets them narrowly to maintain technical control.

These translation gaps are reinforced by commercial and investment pressures. Compliance work in RS development is typically downstream of product strategy and is evaluated through demonstrable artefacts such as audits, documentation, and explainability reports. Venture capital expectations, procurement standards, liability management, and deployment timelines incentivise forms of 'procedural compliance' that minimise disruption to optimisation objectives. In this context, boundary objects do not merely mediate understanding; they become instruments that allow regulatory alignment without requiring changes to revenue models or performance metrics. This dynamic aligns the boundary-object problem with the economic lock-in mechanisms discussed later in this chapter.

Technological solutionism

Technological solutionism refers to the belief that complex social and political problems can be reframed as discrete, technical challenges and resolved through engineering solutions. The term emerged in early 2000s discourse on information and communication technologies for development, where optimism about digital interventions often ignored structural inequality and local context (Heeks, 2002; Toyama, 2011). The idea gained prominence in critical scholarship that examined how technology projects displaced political engagement and institutional reform. While the critique circulated in various

domains, it was popularised by Evgeny Morozov, who documented how private technology firms reframed political crises, public health failures, and educational inequality as problems of information access or interface design (Morozov, 2013). Morozov's contribution was to synthesise these critiques under a single concept and to trace how solutionist logics allowed companies to avoid scrutiny by offering scalable technical responses in place of systemic change. The core mechanism of technological solutionism is reframing (Sætra & Selinger, 2024, 2023). A socially embedded problem is transformed into a system-level optimisation task. For example, public safety is reduced to predictive policing models trained on historical crime data. The algorithm treats biased enforcement patterns as neutral evidence, reinforcing patrol in already patrolled areas and excluding structural variables such as housing instability or racial profiling (Bhuiyan, 2021). Health inequality is reenvisioned as a behavioural issue to be addressed through step counters and gamified fitness apps. The result is a model where well-being becomes a personal data task, not a collective responsibility (Barracough, 2024). Education policy provides a further illustration. Algorithmic grading systems transform pedagogy into a test-score optimisation task, reducing learning to measurable outcomes that ignore context and creativity. Welfare administration follows a similar path. Digital monitoring reframes poverty as a matter of data collection, where compliance is tested through automated checks rather than through structural analysis of inequality. In both cases, systemic issues are translated into datasets and optimisation targets. What began as political or social debates is reframed into engineering routines that produce reports without necessarily driving reform. Environmental degradation is reframed as a problem of waste monitoring, addressed through smart bins and

data dashboards. Users are encouraged to sort, scan, and log waste, while upstream causes such as industrial packaging norms or regulatory failures are ignored (BBC, 2024). Each case shares the same structure: a political or collective issue is abstracted into data, then recomposed as an engineering problem that avoids the original site of conflict. As a result, technological interventions offer the appearance of action without redistributing power or addressing underlying causes. In this framing, responsibility for social outcomes is more frequently articulated in relation to platform design choices, while structural factors such as policy, funding, or regulation receive less discursive attention. This reframing extends to RS development, where criticism of algorithmic bias or democratic erosion is met with technical adjustments that preserve the optimisation objective. For instance, fairness re-ranking layers are introduced to redistribute exposure scores while leaving the loss function unchanged (Singh & Joachims, 2018). In doing so, platforms maintain the same engagement-based ranking logic that generated the initial concern. Similarly, privacy criticisms are addressed through on-device computation or differential privacy noise. These solutions restrict data exposure but increase dependency on proprietary infrastructure. They allow platforms to respond to public demands while maintaining continuity in optimisation logic. The pattern reflects what Bardzell and Skaug Sætra have identified as a tendency in human-computer interaction to treat ethics as an add-on rather than a foundation (Sætra, 2021; Bardzell & Bardzell, 2013). Technological solutionist design allows systems to incorporate critique while structurally resisting transformation. This logic is particularly persistent because it aligns with institutional incentives. Technical solutions are easier to scale, evaluate, and monetise. They can be benchmarked,

versioned, and deployed across use cases. Structural solutions require deliberation, redistribution, and collective decision-making. In the context of RS, a fairness patch can be implemented within the architecture of engagement optimisation. A change to the optimisation objective, however, would alter recommendation logic, business models, and performance targets. Technical solutionism thrives in this gap between symbolic responsiveness and operational continuity. It provides audit artefacts without altering incentives. As Chapters 3 and 4 show, RS interventions that adopt solutionist logic tend to produce narrow forms of accountability. YouTube's response to radicalisation criticism, for example, involved labelling content from 'authoritative sources' while maintaining the same watch-time metric (YouTube Team, 2019; Tufekci, 2018; Ribeiro et al., 2020). This addressed the visibility of legitimacy without questioning the engine that amplifies extreme content. In a privacy context, cookie deprecation led to local inference models being embedded in user devices (Pujol, Hohlfeld & Feldmann, 2015; Papadopoulos, Kourtellis & Markatos, 2019; Papadogiannakis et al., 2021). This limited third-party access, but increased the opacity of on-device RS. Each solution preserves platform control while recoding the critique into a technical requirement. Technological solutionism can be analysed as a pattern in which structural critique is translated into manageable engineering inputs, rather than as a failure to recognise complexity. It displaces contestation into technical domains and translates normative concerns into optimisation problems. In RS design, this produces systems that appear to address bias or risk, but only within the boundaries set by existing architecture and platform goals.

Each technical solution extends platform reach by gathering new behavioural data, while overlooking model scale, the variability of human behaviour, and questions of human agency. Structural inequities remain intact. What appears as meaningful problem-solving often functions instead as market expansion, presented through optimisation rhetoric. Privacy by Design illustrates this logic clearly (Schaar, 2010). Originally conceived as a normative framework for embedding privacy into system architecture, it has been absorbed into a compliance toolset that allows systems to grow without altering core extraction mechanisms (Veale, Binns & Ausloos, 2018; Cohen, 2019). Developers interpret privacy through the lens of data security and policy constraints, not normative deliberation (Hadar et al., 2018). The architectural dimension of Privacy by Design is rarely implemented. Instead, privacy is addressed through encryption, access restrictions, or audit logs, all of which increase the technical surface of the system without limiting its underlying operations. This results in technological solutionism, where social and political problems are transformed into engineering tasks, making critique actionable without making it transformative. Illich and Lang warned that industrial remedies often amplify the dependence they claim to address (Illich & Lang, 1973). Privacy by Design does not reduce the platform's reach or the user's exposure. It shifts the location of computation, embeds control deeper into proprietary infrastructure, and offers interface-level controls that simulate agency. The user sees toggles and consent flows, but the system remains designed around capture, retention, and inference. Other regulatory frameworks show the same absorption. Safety by Design in aviation became codified into checklists that certified compliance without revisiting organisational priorities (van Gelder et al., 2021; Gudmundsson, 2013; Denney, Pai &

Whiteside, 2019). Similarly, Security by Design in software engineering reduced broad concerns about user vulnerability to cryptographic standards and penetration tests. In each case, normative frameworks were translated into procedural templates that maintained system continuity. Privacy by Design follows this trajectory. Its critical edge is softened as it becomes a compliance instrument rather than a structural limit on surveillance. Professionalised medicine, as Illich argued, expanded institutional authority under the guise of care. Privacy by Design performs a similar function. It institutionalises privacy as a measurable property to be certified, rather than a right to be upheld. This transition enables companies to treat privacy as a feature rather than a constraint. Morozov identifies this manoeuvre as a central move of technological solutionist thinking (Morozov, 2013). Each social issue becomes a design prompt. Privacy becomes a matter of system configuration. Fairness becomes a re-ranking task. Transparency becomes a logging feature. These interventions generate documentation and satisfy auditors, but do not challenge business models or power asymmetries. As Privacy by Design is reinterpreted by developers and organisations, its critical edge is weakened. The principle of embedding privacy is retained, but its meaning shifts from limiting what the system can do to proving that the system has responded. What is embedded is not resistance but flexibility. In this form, Privacy by Design reinforces the very system it was meant to constrain. Privacy is no longer a boundary on data collection, but a new layer of technical work that deepens the platform's infrastructural footprint.

RS illustrate this cycle. Early critiques of RS focused on filter bubbles, ideological polarisation, and demographic bias. These concerns pointed to structural problems in how RS architectures prioritise engagement, suppress minority viewpoints, and reproduce existing hierarchies. In response, the field introduced fairness re-ranking layers. These post-processing modules adjust exposure scores across demographic categories to produce statistically balanced outputs. However, such interventions do not alter the objective function that governs model behaviour. The underlying loss function is still tied to watch time, click-through rate, or retention. The system continues to optimise for engagement, only now with fairness constraints applied at the output stage. This architecture reflects the logic of technological solutionism. Critique is absorbed as a discrete engineering task that does not question the premises of the system. Fairness becomes quantifiable and therefore implementable within the existing optimisation frame. It is addressed without revisiting broader questions about power, representation, or epistemic harm. When fairness layers succeed, platforms gain compliance legitimacy. At the same time, they maintain the same incentive structure that prioritises profitable, popular, or ideologically aligned content. The RS still filters and ranks based on behavioural signals, not civic relevance or deliberative diversity. This pattern mirrors the trajectory of Privacy by Design. Rather than redesigning systems to restrict capture or profiling, platforms implement privacy as a set of technical features. These include logs, settings, or local processing modules that do not limit the scope of data collection. Fairness layers operate in a similar way. They add a visible mechanism of correction without interrupting the upstream logic. In both cases, the intervention validates the system's output while preserving the structure that generated the

critique. The RS appears to respond to public and regulatory pressure, but continues to sort and amplify content in line with institutional goals. Technical solutions demonstrate accountability through statistical evidence and documentation, but the function of the system remains unchanged.

As shown in Chapter 3, YouTube's response to criticism over political radicalisation follows the same solutionist logic. The platform introduced authoritative-source labels on election-related content, presenting the change as a transparency measure. However, the core optimisation objective of maximising watch time remained unchanged. The system continued to amplify emotionally charged and partisan content, including videos with high engagement but reduced linguistic variation. In the Democratic party discourse case, lexical diversity dropped during politically sensitive periods, particularly in January 2024, while negative sentiment increased due to the January 6th discourse (BBC, 2021). This suggests that the RS continued to reward content that reinforced ideological uniformity and affective intensity, regardless of surface-level interventions. The addition of authoritative labels did not modify the ranking logic. YouTube's recommendation model prioritises engagement signals such as session duration and click-through rate (Covington, Adams & Sargin, 2016). The labels operate as an overlay. They do not interact with the model's internal objective. Content flagged as authoritative is still subject to the same ranking mechanisms as other videos. In practice, the most engaging material continues to dominate exposure. The intervention adds a layer of perceived neutrality but leaves the system's behavioural incentives intact. At the same time, the platform restricted access to raw recommendation data. Researchers and auditors were

redirected to platform-managed dashboards and curated summaries, reflecting a broader trend where companies such as Spotify have begun restricting API access and its functionality (Dees, 2024). These replacements limited independent evaluation and reduced transparency. While presented as a privacy or safety measure, the restriction increased the platform's control over the research process. External actors could no longer verify how content was ranked or distributed. This shift can be read as concentrating epistemic authority within the platform while preserving the appearance of openness. These developments reflect a broader pattern in RS design. Critique is acknowledged, translated into a technical feature, and absorbed without altering system logic. Authoritative labels, such as fairness layers or privacy settings, become compliance artefacts. They signal responsiveness to regulation or public pressure while allowing the underlying optimisation process to continue. Engagement remains the primary value. Adjustments occur at the interface, not in the model's architecture. YouTube's response illustrates how platforms accommodate critique through modular interventions that protect existing priorities. Reform is simulated through visibility controls rather than structural change. Authoritative labels do not alter what the system values or how it selects content. They communicate credibility while the model continues to promote patterns of polarisation and repetition. The result is a consolidation of control over knowledge distribution, achieved through superficial corrections that maintain the system's operational logic.

Privacy concerns triggered a similar shift. As cookie tracking faced regulatory and public scrutiny, platforms introduced on-device RS and differential privacy mechanisms. These approaches limit raw data transmission, which

reduces third-party tracking. However, they require deeper integration of proprietary models directly into user devices (Mayer & Mitchell, 2012; Englehardt et al., 2015; Clarke, 2024; Chen et al., 2021; Atrey et al., 2023b, 2023a). This change decreases observable data flow while increasing reliance on opaque local inference. The system becomes less auditable from outside. The result is not greater user control but a redistribution of informational asymmetry. Bad actors no longer need to intercept network traffic to infer user behaviour. Instead, they focus on model reconstruction, a process where adversaries query deployed models to extract training data characteristics or infer sensitive attributes (Zhao, Alwidian & Mahmoud, 2022; Chakraborty et al., 2018; Wu & Zhou, 2021; Lumbut & Ponnoprat, 2025; He, Zhang & Lee, 2019). Multiple studies have demonstrated that even models designed for differential privacy remain vulnerable to inversion attacks and membership inference when access is sustained over time. Carlini show that language models trained on public and private data leak identifiable sequences, even when sampling is restricted (Pintor et al., 2022; Carlini, Feldman & Nasr, 2022; Nasr et al., 2021; Carlini et al., 2021). Fredrikson, Jha and Ristenpart demonstrate that facial recognition models can be queried to reconstruct original training images (Fredrikson, Jha & Ristenpart, 2015). These findings show that the threat has shifted from data transport to model exposure. The platform appears more secure, but model-level privacy risks persist. Each patch extends algorithmic control into new domains. Sentiment analysis, originally developed to optimise advertising placement, now underpins automated labelling systems for child-safety compliance (Qiu et al., 2010; Sánchez-Núñez et al., 2020). These systems classify emotional tone in videos to flag potentially harmful content. In practice, this leads to false positives

that disproportionately affect marginal voices, especially those using non-standard linguistic features or emotionally coded discourse. Geo-context filters, first introduced to suppress location-based extremist content, have since been repurposed for price discrimination (Azzolina et al., 2021; Hannak et al., 2014; Faisal, 2023). Studies by the Norwegian Consumer Council and academic audits of travel platforms have shown that users in different locations receive different price options based on geolocation inference (Kaldestad, 2024; Chowdhury & Stoica, 2012). These examples show that mitigation tools often migrate. A system developed to meet one regulatory requirement becomes infrastructure for an unrelated commercial aim. Technical adaptability allows the same mechanism to address multiple goals without renewed scrutiny or ethical evaluation. Technological solutionism persists because success is measured narrowly. Fairness is validated through performance on balanced test sets, not by long-term impact on discourse. Privacy is confirmed through reduced re-identification rates, not by examining shifts in control or access. Metrics focus on local optimisation. Broader effects such as semantic narrowing, content homogenisation, and the consolidation of surveillance capacity remain outside the scope of formal evaluation. The model appears compliant. The structure remains unchanged. In the European context, the Digital Services Act, adopted after the empirical phase of this research, introduces new data-access and transparency provisions that may partially counter this trend by expanding formal opportunities for independent researcher auditing of recommender systems. Whether these mechanisms meaningfully increase external scrutiny of optimisation practices, or instead channel research through platform-mediated

interfaces, remains an open empirical question and a priority for future investigation.

Economic incentive lock-in

Noble traced the manner in which numerically controlled machine tools emerged under military sponsorship, embedding capital's priorities of speed and interchangeable labour into production hardware (Noble, 1978). Once installed, the equipment redefined notions of 'efficiency' and 'skill,' allowing management to frame workforce displacement as a technological inevitability rather than a strategic choice. Investment moved into tooling, software, and training created lock-in, as alternative shop-floor models became unviable without access to the committed capital. What began as technical modernisation quickly became institutional dependence, as infrastructures and labour routines were reshaped to fit the logic of automated control. Lock-in, in this context, refers not only to technical inflexibility but to structural reinforcement of one model of production. The equipment did not simply automate tasks. It reorganised decision-making around the machine's logic. Skills grounded in judgement and adaptation were displaced by procedural compatibility. Once operational systems adjusted to this framework, deviation became costly, not because alternatives were ineffective, but because they no longer fit the technical or organisational frame. As routines, training, and investment aligned around the installed model, exit options narrowed, and the system acquired momentum through its own dependencies. This form of technological persistence makes alternative configurations appear inefficient or impractical, even if they offer social or epistemic advantages. Efficiency becomes redefined by the capabilities of the infrastructure. Design

decisions are driven more by what systems can work together than by thoughtful discussion. Lock-in thus operates through cumulative reinforcement, where each new component assumes the validity of earlier decisions. Organisational stability, supply chain optimisation, and investment planning all align with the same technical baseline. Reversal would require a parallel redesign of multiple layers, from hardware to workflow to institutional justification. This historical pattern provides a foundation for analysing contemporary RS infrastructure. RS architectures are shaped by capital-intensive choices that produce long-term dependency. Once performance metrics are embedded into training routines, evaluation practices, and interface logic, the cost of revision rises. Optimisation no longer applies to content or users alone, but to the internal structure of the system itself. This entrenchment is not the result of explicit resistance to change, but of the cumulative dependencies that make change structurally and economically unattractive.

Economic-incentive lock-in, as formalised by David, describes how early technological choices become self-reinforcing over time, even when objectively better alternatives emerge (David, 1985). David's framework builds on the concept of path dependence but introduces a specific economic logic to explain how and why certain systems persist. He argues that initial adoption, followed by increasing returns to scale, network effects, and the accumulation of complementary investments, locks actors into suboptimal configurations. The QWERTY typewriter keyboard layout serves as his case study. Although more efficient layouts such as Dvorak existed and had demonstrable performance advantages, QWERTY survived because typists, training schools,

manufacturers, and software systems had already standardised around it. The growing installed base created coordination benefits that outweighed the technical merits of switching. As adoption increased, switching costs rose, not just in terms of retraining, but also in lost interoperability, reduced productivity during the transition, and the disruption of shared routines. Other media technologies illustrate the same outcome. The VHS format dominated over Betamax despite lower technical fidelity because it aligned with rental-market strategies and manufacturer agreements (Cusumano, Mylonadis & Rosenbloom, 1992). GSM became the global mobile standard while technically competitive alternatives such as CDMA persisted only in limited regions (Venkatesan, 2013). These cases show how early coordination advantages outweigh technical considerations. RS metrics display the same dynamic, where adoption and standardisation create persistence independent of epistemic outcomes. David's argument is that economic rationality in the short term can produce systemic irrationality in the long term (David, 1985). Decision-makers optimise for compatibility, not performance. What was initially a contingent choice becomes embedded through infrastructure, institutional inertia, and market expectations. This lock-in dynamic can be analysed as a consequence of rational actors adapting to existing conditions, rather than as a simple flaw or market failure. Once the cost of deviation exceeds the perceived benefits of change, even inefficient systems endure. Efficiency, in this framework, is not judged against ideal benchmarks but against the structure of sunk costs and mutual dependencies. This is comparable to how a hospital interprets blood test results not against universal clinical benchmarks but relative to reference intervals 'reported by the manufacturer or as established by another laboratory' (Katayev,

Balciza & Seccombe, 2010). This model offers a precise language for understanding how digital infrastructures, including RS, evolve. The adoption of a specific optimisation metric, such as click-through rate or watch-time, mirrors the early adoption of QWERTY. Once the metric is embedded into machine learning pipelines, engineering workflows, advertiser contracts, and performance reviews, it becomes the default measure of system success. As more actors coordinate their activities around this metric, its authority becomes self-reinforcing. Platforms calibrate ranking models to it. Creators adjust production strategies to optimise for it. Regulators audit platform behaviour based on its outputs. Switching to a different optimisation objective would require reconfiguring not only technical systems but also organisational incentives and stakeholder expectations. The QWERTY framework also helps clarify why interventions aimed at mitigating algorithmic harm often leave the core optimisation structure untouched. Mitigation layers, such as fairness adjustments or labelling mechanisms, are applied downstream, after the initial metric has already shaped content exposure. These interventions preserve the locked-in metric while addressing its side effects. As David showed, once a technical path accumulates economic and institutional weight, deviation requires more than evidence of inefficiency. It requires disrupting the network of aligned interests that maintain it. In the case of RS, the lock-in is reinforced not only by platform architecture but by the monetisation models, regulatory assumptions, and cultural expectations that have grown around engagement as a proxy for value. David's analysis therefore moves the discussion from whether a system works well to why it persists despite critique, and what forms of structural dependence sustain its continuity.

Engagement-optimised RS follow a similar economic trajectory to that described by Noble and David. Their development is shaped by the need to generate predictable returns through measurable user activity. Advertising remains the primary source of revenue for most large platforms, which creates a structural requirement to demonstrate attention in quantifiable form. Metrics such as click-through rate, watch-time, scroll depth, and session duration have become the standard proxies for value. These indicators do not measure user satisfaction, understanding, or long-term civic benefit. They measure the frequency and duration of interaction in formats that can be sold to advertisers. The stronger the metric signal, the higher the potential for monetisation through targeted placements, real-time auctions, and predictive exposure strategies. To support this model, platforms invest heavily in the technical infrastructure that sustains metric collection and amplification. This includes data centres capable of storing behavioural traces at scale, feature-store architectures that aggregate contextual signals, and A/B testing frameworks that continuously refine model performance against engagement indicators. Engineering priorities shift accordingly. Success is defined by model performance on business-aligned benchmarks, such as reducing bounce rate or increasing click conversion. These feedback loops incentivise fast-cycle content evaluation, favouring formats that produce immediate response. Videos that trigger comments or shares within the first thirty seconds, headlines that generate impulse clicks, and posts that provoke polarised reactions are all privileged by ranking models trained on engagement objectives. This infrastructure hardens the logic of short-term optimisation into system architecture. Content that does not produce rapid or

repeatable feedback is filtered out early. Educational material that requires sustained attention, minority perspectives with limited prior engagement, or deliberative formats that resist emotional triggers are algorithmically downranked. The system's design appears data-driven and agnostic, but in practice it encodes a preference for content that maximises measurable output under time constraints. As in Noble's research, where industrial tools displaced collaborative labour under the banner of technical advancement, RS rationalise the marginalisation of alternative content forms as a natural result of user behaviour. The fact that content does not perform under engagement metrics is treated as evidence of irrelevance, not as a signal of metric bias. Examples of this logic are visible across platform types. On YouTube, recommendation algorithms prioritise watch-time, leading to the dominance of long-form commentary videos that sustain user attention through sustained conflict or speculation. On Instagram and TikTok, content that generates fast responses, such as challenges or reactions, is favoured over slower aesthetic or narrative development. News platforms that integrate RS features see rapid traffic growth for emotionally charged or misleading headlines, as click-through is rewarded even when comprehension or accuracy declines (Feng et al., 2020; Raza & Ding, 2022; Mizgajski & Morzy, 2019). Over time, these performance patterns feedback into content production, as creators, influencers, and publishers adapt their strategies to fit the metric logic. The result can be read as a shift in how digital communication is organised, with metrics aligned to advertising imperatives becoming more salient than considerations of epistemic diversity or democratic value. These recursive dynamics transform creative labour. Content producers on YouTube adjust thumbnails, titles, and posting schedules to align

with algorithmic cues. TikTok creators adopt short-form strategies optimised for immediate reactions because those formats secure visibility. The labour process becomes organised around platform metrics rather than around communicative goals. This resembles Braverman's account of industrial deskilling (Elger, 2024, 1979), where workers adapt to machine rhythms rather than exercising judgement. In RS environments, epistemic agency is displaced by compatibility with engagement metrics. Creativity becomes a response to optimisation rather than a driver of discourse.

Lock-in operates through the accumulation of complementary assets that align incentives across different actors within the RS ecosystem. These assets include contractual arrangements, behavioural expectations, and technical dependencies that collectively reinforce the dominance of engagement-based metrics. Brand advertisers, for instance, do not buy visibility in the abstract. They purchase targeted impressions that are calculated and delivered by the RS according to performance benchmarks such as click-through rate or audience segmentation depth. The ranking engine packages user attention into monetisable units, which are then priced and sold in real time through programmatic advertising systems (Zhang, Yuan & Wang, 2014; Yuan, Wang & Zhao, 2013). This pricing model depends on stable and predictable engagement metrics, which in turn shape how RS are trained and evaluated. Content creators, facing visibility thresholds that determine monetisation and reach, adapt their production strategies to align with the algorithmic cues. Thumbnail selection, posting frequency, format duration, and topical framing are all adjusted to fit engagement optimisation criteria. This recursive adaptation deepens the

system's alignment with its original metric, as successful content further entrenches the performance standards that define visibility. Over time, the platform's content ecology is structured not by expressive diversity but by compatibility with the RS logic. The metric becomes the organising principle of creativity, narrowing the conditions under which content is judged as successful. Venture capital valuations also depend on the ability to demonstrate continuous user growth, platform retention, and engagement depth (Hall & Hofer, 1993; Montanaro, Croce & Ughetto, 2024). These indicators are treated as proxies for future profitability and competitive strength. Start-ups and platforms seeking funding optimise for the same engagement metrics because they offer investors comparable performance indicators across portfolio companies. As a result, any deviation from engagement-optimised strategies threatens not only technical consistency but financial valuation and investor confidence. Together, these interdependencies raise the cost of transition. Shifting toward slower, deliberative RS models would require advertisers to accept new metrics, creators to change production norms, and investors to reassess value criteria. The entrenchment of complementary assets ensures that the metric is not just a technical tool but a shared reference point that stabilises a broader economic order.

When regulation intervenes, it confronts an ecosystem whose operational logic and revenue structure depend on maintaining the existing loss function. Optimisation metrics are not peripheral components but central elements that align technical infrastructure with commercial or political goals. These metrics determine what content is surfaced, what counts as successful interaction, and how system performance is evaluated. Attempts to introduce oversight or

constraints often target outputs rather than the core mechanism that generates them. As a result, most interventions focus on visibility, explainability, or procedural fairness without questioning the metric that structures decision-making throughout the RS pipeline. Governance frameworks tend to prioritise transparency, user control, and risk mitigation. In practice, this is implemented through reporting interfaces, summary statistics, and documentation of system behaviour. However, the data disclosed typically reflects aggregate measures and excludes the underlying optimisation rationale. Platforms publish figures on engagement levels, user reports, or the share of recommended content that meets specific criteria. These statistics give the appearance of openness while preserving the internal logic of engagement-based design. As long as the loss function continues to prioritise metrics such as watch-time or click-through, the structure of content filtering and ranking remains unchanged. Outputs may shift slightly, but the incentive to produce measurable interaction persists. Chapter 3 identified declining linguistic diversity in engagement-optimised environments, suggesting a tendency toward echo chambers and reduced exposure to alternative perspectives. Content with dominant lexical patterns was surfaced more frequently, while variation diminished across recommendations. This was not due to explicit exclusion, but a consequence of performance metrics that prioritised engagement consistency. Chapter 4 focused on the institutional framing of AI governing frameworks in the EU, USA, and Russia. It showed that EU and US draft legislation prioritised transparency, risk classification, and auditability, but avoided mandating structural changes to platform objectives or optimisation frameworks. Regulatory attention concentrated on procedural compliance rather than on the epistemic implications of RS design. In Russia,

political mandates replaced commercial imperatives, but the metric-based optimisation logic persisted. Relevance was redefined to prioritise ideological coherence, yet the technical architecture of RSs remained shaped by institutional power through performance-aligned parameter. In this setting, the ranking model did not reward interaction volume but narrative conformity. The result mirrored commercial systems in form even as it diverged in substance. Economic-incentive lock-in explains the durability of these configurations. RS models are embedded in layers of technical and organisational interdependence. Training procedures, performance benchmarks, user modelling systems, and content pipelines are calibrated around specific loss functions. Any significant change would require redesigning workflows, rewriting interface logic, and revising stakeholder contracts. Platforms have invested heavily in tooling, datasets, infrastructure, and developer expertise that reinforce the current metrics. These sunk costs make structural revision costly and time intensive. Changing the objective function would undermine the value of existing assets and destabilise internal coordination. Regulatory interventions often face limits in authority or access when attempting to challenge these embedded structures. Audits focus on content categories, output proportions, or compliance with transparency requirements. They rarely extend to the assumptions that govern model design or the objectives that drive optimisation. Platforms retain control over how performance is defined and how system behaviour is reported. Regulators respond to what is made visible, which limits their ability to address deeper forms of epistemic exclusion or structural bias. Even when external pressure mounts, platforms can demonstrate progress through modifications to interface elements, labelling systems, or procedural safeguards. These changes are presented as

evidence of adaptation, yet the ranking logic remains intact. As Noble observed in industrial contexts, technical systems acquire momentum through infrastructure, investment, and standardisation. Once installed, a design choice becomes naturalised. It is treated not as a decision but as a baseline. Within RS environments, the economic and technical commitments to engagement-based metrics perform the same function. They transform an initial optimisation preference into a foundational constraint. Structural alternatives can appear impractical, not necessarily because they are unworkable, but because they do not align with the existing logic of the system. This is why technical mitigation and legal intervention often stabilise the system rather than alter its direction.

Conclusion

RS are not neutral tools but cultural technologies of omission, shaped by historical, technical, and economic legacies that constrain epistemic agency. This chapter has shown how path-dependence locks in early design choices, how instrumental rationality prioritises efficiency over diversity, and how boundary-object translation produces governance gaps. Technological solutionism adds layers of complexity without addressing root causes, while economic lock-in solidifies designs that favour engagement or control over pluralism. Across both English market-driven and Russian state-directed contexts, RS converge on curated absence, highlighting content that aligns with prior commitments of profit or power while sidelining broader discourse. These self-reinforcing patterns reveal RS as heirs to a lineage of knowledge-filtering tools, from catalogues to broadcast schedules. The consequences extend to epistemic agency. Citizens engage with information environments shaped by

omissions that are structural rather than incidental. The narrowing of linguistic diversity observed in Chapter 3 illustrates this effect. Declines in lexical variation limited the range of perspectives available for deliberation. When RS amplify only those signals that align with pre-encoded objectives, the possibility of informed choice diminishes. Agency becomes reactive within a pre-filtered space, rather than deliberative across a plurality of views. This confirms that RS are not just technical infrastructures but epistemic environments that condition participation. Governance struggles to catch up because mismatched vocabularies and incentives obscure structural biases. The contribution of this chapter is therefore not to establish a causal relationship between RS and democratic decline, but to theorise why empirically observed patterns of linguistic narrowing, regulatory misalignment, and research prioritisation recur across contexts. The theories deployed here explain persistence, translation, and reinforcement, not direct behavioural causation.

The individual concepts in Fig 28 are not new. Path dependence, instrumental rationality, boundary objects, technological solutionism, and lock-in are established frameworks in STS, sociology, and political economy. The contribution of Fig 28 is a synthetic chaining that explains how technical choices become durable through a sequence of translation and reinforcement. The proposed chain is intended as an organising mechanism for interpreting RS governance across chapters, rather than as a claim of a novel theory family. In the European context, the cycle in Fig 28 can be read against the AI Act's built-in learning mechanisms, which operate mainly through post market monitoring and incident feedback. For high-risk systems, providers must maintain post

market monitoring and feed observations into compliance documentation, and they must report serious incidents to market surveillance authorities, creating a channel for iterative correction over time. In addition, regulatory sandboxes and evolving codes of practice, overseen through EU level coordination, can function as upstream learning spaces where requirements are interpreted and operationalised before full deployment. However, these loops may only partially disrupt the cycle described in Fig 28 because they tend to learn from observable harms and compliance artefacts, while the boundary object problem and economic lock-in can leave optimisation objectives and incentive structures intact.

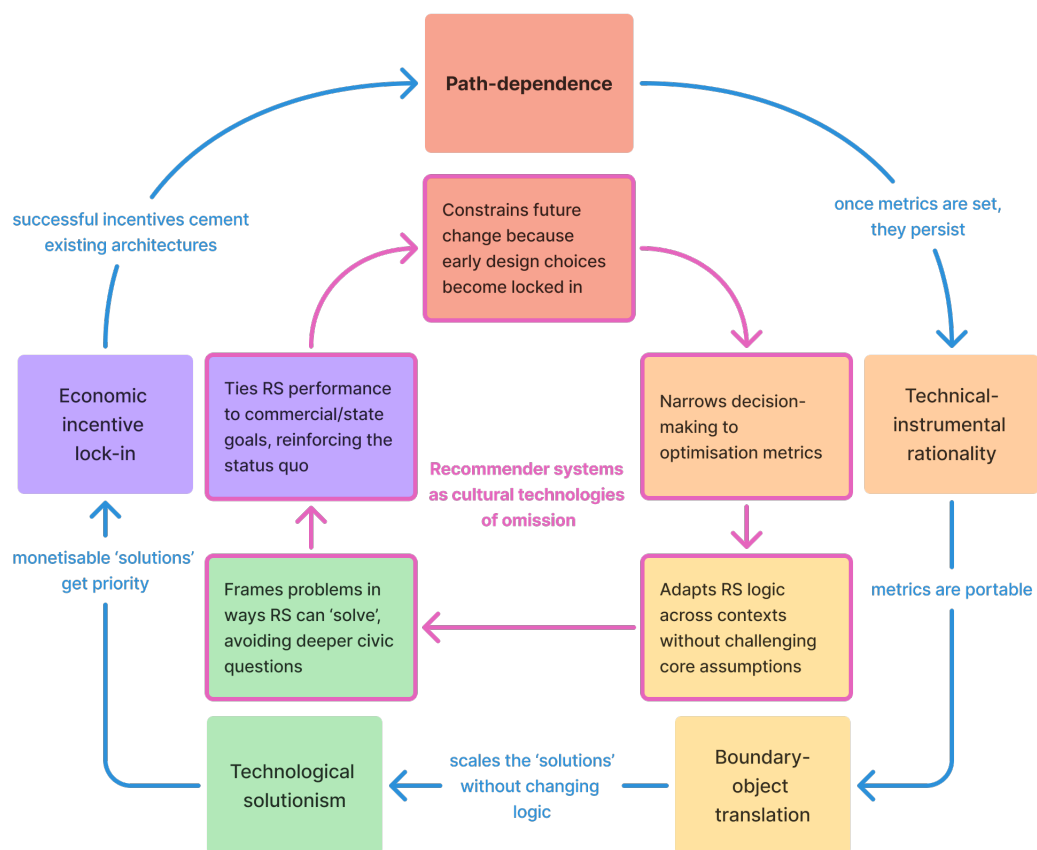


Figure 30 Reinforcing cycle of RS epistemic control.

Chapter 7: Conclusion

This thesis investigates how algorithmic infrastructures shape public access to information relevant for civic participation, focusing on the constraints imposed by recommender systems (RS) on epistemic agency. Across digital platforms, RS filter and prioritise content based on engagement-led metrics such as watch-time or click-through rate. These outputs are not neutral. They constitute a layer of mediation that shapes visibility, limits discursive plurality, and conditions the terms under which knowledge circulates in public life. The capacity to participate in cultural and political discourse is no longer tied just to the availability of information. It also depends on whether access routes present information in ways that allow citizens to use it meaningfully in public debate, for example by ensuring that content is surfaced in a timely, comprehensible and contextually relevant form. RS influence this shift through their operational role in determining the visibility and relevance of civic participation. Rather than offering transparent access to diverse perspectives, RS amplify patterns that reflect platform-specific optimisation strategies. In multilingual and politically polarised environments, this results in asymmetries of representation and visibility, privileging dominant-language and majority-aligned narratives while marginalising minority-language and dissenting perspectives. RS do not just mediate content. They co-produce discourse, prioritising certain framings, actors, and terms over others. These patterns are analytically relevant to which narratives become visible, which remain peripheral, and which receive less discursive attention within RS-mediated environments. This thesis centres on the concept of epistemic agency and its role in democracy, defined here as the capacity of users to influence and shape shared knowledge systems through

active participation in civic discourse. Epistemic agency is not evenly distributed. It is contingent on linguistic access, cultural framing, and the infrastructural design of communication systems. Language is not just a medium. It defines how issues are framed, what counts as evidence, and who is considered a credible speaker. The thesis connects epistemic agency to the capacity to act through language in civic contexts, arguing that algorithmic filtering, whether intentional or incidental, limits not only what is seen but also which linguistic forms enter circulation. This matters because limiting available forms constrains how issues can be framed, which affects the scope of reasoning and the range of perspectives that can participate in public discourse. As certain content is deprioritised, the associated vocabulary and discursive variety are likely to be suppressed, narrowing what can be said and by whom. Prior work has not addressed this relationship adequately, as much of the RS literature operates within English-only corpora and datasets, assuming language neutrality while relying on models trained predominantly on monolingual input. User-centred studies often emphasise perceived fairness or usability, without situating RS within broader epistemological or sociotechnical frameworks. Empirical investigations tend to isolate platforms, national, or cultural contexts, with limited attention to how governance, infrastructure, and language interact at scale. The result is a body of research that often focuses on isolated effects, with more limited attention to the structural logics that shape epistemic participation. This thesis responds with a cross-platform, cross-linguistic, and epistemology-focused design. It combines empirical data from YouTube and Dzen with comparative analysis of AI governance frameworks in the EU, USA, and Russia, and a critical analysis of the terminology used in RS research discourse in

English and Russian. This approach enables analysis across different levels of how RS function not just as recommendation tools but as epistemic infrastructures that shape discursive possibilities, showing how language both shapes and is shaped by the technological and cultural affordances that define each context. By focusing on linguistic variance and platform logic, the thesis presents RS as civic technologies with epistemological consequences. It examines how specific RS algorithmic practices such as ranking, filtering and optimisation for engagement metrics, and the language used to describe and regulate them, determine which knowledge becomes visible and which is suppressed.

According to this perspective, a RS is not simply recommending. It structures participation by filtering the conditions under which meaning is produced, distributed, and acted upon in democratic discourse. This framing contrasts with most existing literature, where RS are examined as technical systems built to solve discrete problems of personalisation, user retention, or content ranking. Across RS design, AI ethics, democratic theory, and digital governance, research develops within disciplinary silos and, beyond this, abstracts these systems from their operational context. RS are analysed as self-contained optimisation engines rather than embedded infrastructures that influence cultural visibility and discursive access. Across these domains, three assumptions recur. First, RS are evaluated primarily through metrics that prioritise efficiency and engagement. Standard measures such as Root Mean Square Error, F1-score, or user click-through rate are designed to maximise system responsiveness, not informational or linguistic diversity. These metrics

encode a narrow view of success where predictive precision is favoured over civic relevance. Second, models are often treated as language neutral. Most RS research uses English-language datasets and assumes linguistic transferability without assessing how conceptual drift or term weighting might shift across languages. This reinforces the idea that linguistic and cultural differences are peripheral rather than integral to system behaviour. Third, users are positioned as passive endpoints. Engagement is measured, optimised, and retained but rarely interpreted as a function of knowledge-making or epistemic intent. There are also clear absences, such as platform-specific logic. While evaluation metrics are discussed in detail, little attention is paid to how they are selected or why certain forms of content are rewarded by loss functions or filtering parameters. As shown in Chapter 2, the choice of training strategy, particularly the loss function, determines which data are reinforced and which are excluded before any recommendation is made. This shapes not only outputs but also the knowledge space in which users operate. While concerns around translation quality or input bias are mentioned, the deeper role of linguistic framing in constructing knowledge categories, authority claims, or civic narratives is left untouched. Research into governance often reduces regulation to a question of oversight, without considering how legal frameworks reflect and sustain platform epistemologies and the value systems that underpin them. The thesis builds on RS design research but shifts the focus from optimisation to epistemic consequence. More accurate recommendations do not necessarily result in better informed citizens. This dissertation examines how accuracy is defined, by whom, and for what purpose. In doing so, it connects platform metrics, legal framings, and research discourse to the broader question of epistemic agency:

whose knowledge and perspectives shape the knowledge environment, and whose are made visible by design.

Chapter 3 observes how RS structure civic access to information across two platforms, YouTube and Dzen. The study used 60 sock-puppet accounts, calibrated by gender, age, language, and behavioural pattern, to simulate users navigating each platform in the lead-up to national elections in the USA and Russia. The accounts were active for four months, enabling observation of recommendation trajectories and the role of language and platform-specific RS logic in shaping surfaced content. This design made it possible to trace recommendation trajectories and isolate the role of language and platform-specific RS logic in shaping what information is surfaced and prioritised. The findings reveal consistent asymmetries in the output of both systems. On YouTube, English-language accounts aligned with Democratic Party discourse (USB) were exposed to lower lexical diversity and greater sentiment volatility compared to Republican-aligned accounts (USR), despite similar vocabulary sizes. This pattern indicates that USB recommendations were more repetitive and displayed greater sentiment volatility, which concentrated exposure on emotionally charged narratives and corresponded to a narrower range of perspectives within the recommendation stream. In contrast, USR displayed more consistent sentiment patterns and broader lexical variation, suggesting a recommendation loop structured around stable ideological content with higher engagement retention. In the Russian context, the asymmetry was more pronounced and consistently observable across recommendation outputs. Dzen (RUR), which delivered pro-war content, exhibited higher lexical diversity and

more dynamic sentiment shifts, reflecting an adaptive strategy aimed at sustaining engagement through repetition of ideologically aligned narratives. YouTube (RUB), the only platform that surfaced anti-war content, showed constrained lexical variation and static recommendation patterns, reinforcing a narrow discursive space. Gender segmentation revealed further disparities. RUR excluded female-specific targeting altogether, while RUB presented a more generalised demographic spread but still prioritised content that aligned with dominant themes. In both cases, RS operated not merely as content filters but as infrastructures of linguistic reinforcement. Content that aligned with prior high engagement narratives, whether mainstream, polarised, or state aligned, was repeatedly surfaced, while alternative or low frequency perspectives were algorithmically marginalised. Language functioned as a key axis of differentiation, with linguistic diversity directly linked to the platform's tuning for retention and ideological conformity. These dynamics shaped not only what users encountered but also how civic discourse was structured, segmented, and stabilised over time.

Chapter 4 examines how RS are addressed within three AI regulatory frameworks: the EU, USA, and Russia. The EU AI Act introduces a risk-based framework, but RS are not explicitly classified as high-risk systems. They are mentioned in provisions on transparency for AI systems that interact with users, particularly in contexts such as content ranking and recommendation. These obligations require that users are informed when they are interacting with RS, but the regulation does not assess how RS shape knowledge access or influence public discourse. References to fundamental rights serve a justificatory function

but are proceduralised through technical compliance mechanisms such as risk documentation and oversight protocols. Multilingualism is acknowledged in implementation, but not as a structural factor in epistemic access. As a result, the regulation imposes limited obligations on RS, with their epistemic implications receiving little explicit consideration. In the USA, the regulatory approach relevant to RS is decentralised and fragmented. There is no binding federal regulation that addresses RS directly: instead, existing policies promote innovation and market competitiveness while encouraging voluntary ethical standards. Regulatory oversight is sector-specific and often non-binding. RS are framed primarily as tools for personalisation and efficiency, not as systems that shape civic knowledge. Language and epistemic access are not considered within these frameworks, and the burden of responsible design rests with industry actors. In Russia, RS are not governed through targeted legislation but are embedded in national strategies that frame digital systems as instruments of technological sovereignty and administrative coordination. Policy documents emphasise alignment with state priorities and centralised content governance, without reference to civic participation. RS operate within a broader information control environment, where their function is shaped by institutional objectives rather than epistemic openness. Across all three contexts, RS are treated as technical tools rather than epistemic infrastructures. Legal definitions of harm focus on measurable risks such as bias or privacy violations, without addressing structural exclusions from discourse. Language is treated as a delivery mechanism, not as a condition of agency. None of the frameworks interrogate how RS filter linguistic variation or delimit access to civic knowledge. This regulatory gap can be read as aligning with optimisation logics and as shaping

the conditions under which epistemic agency is articulated within RS governance.

Chapter 5 examines the linguistic and conceptual framing of RS discourse across two corpora reflecting the discourses not of users or regulators, but the creators of RS: English-language publications accessed through OpenAlex and Russian-language publications sourced from eLibrary.ru. The analysis draws on the IEEE Ethically Aligned Design glossary to identify key terms and conceptual framings that structure RS discourse, focusing on how selected terms appear, shift, or disappear across the English and Russian corpora. Rather than treating terminology as neutral, the chapter analyses these terms as examples of epistemic encoding in which priorities, assumptions, and institutional alignments are embedded. The aim is to show that research itself acts as a discursive infrastructure that shapes which framings of RS become standardised, and which remain marginal or unrecognised. In the English-language corpus, RS discourse is dominated by performance metrics, ethical formalism, and standardised technical vocabularies. Terms such as ‘autonomy’ are used to describe system functionality rather than user capacity, and ‘accountability’ is often framed as model traceability or developer responsibility rather than civic consequence. These terms circulate within a formalised technical discourse that emphasises stability, reproducibility, and standardised interpretation. This discourse aligns with Western regulatory interests and technological paradigms, positioning RS as optimisable tools within a market of ethical trade-offs. The language is formal, bounded, and repeatable, designed to support reproducibility and integration into technical systems. This discourse privileges predictability over debate and tends

to situate epistemic agency outside the explicit scope of system design. In the Russian corpus, the vocabulary reflects a different set of constraints. While technical terms overlap with the English discourse, ethical and civic framings are limited or absent. Concepts such as 'autonomy' are rarely used outside the context of system independence, and terms such as 'bias' or 'accountability' appear infrequently or are subsumed under broader state-aligned goals such as national security or cultural stability. The lack of variation in terminology reflects both institutional alignment and restricted access to diverse scholarly sources. The discursive field is narrower, with fewer references to critical or international literature and a stronger emphasis on implementation over critique. This constrains the epistemic range of the research and reinforces the framing of RS as instruments of administrative or commercial efficiency. These differences are not stylistic. They represent distinct regimes of knowledge production that shape how RS are understood, debated, and developed. In both corpora, language encodes the epistemic boundaries of what RS are for, who they serve, and what constitutes a legitimate concern. In the English case, the dominance of modular ethics and audit-centric terms marginalises relational and linguistic perspectives. In the Russian case, the compression of discursive space around state narratives excludes critical framings and constrains pluralism. Both contribute to the civic empowerment gap identified earlier: one through dominance, the other through restriction. The consequence is a narrowing of epistemic agency in both directions. The English-language corpus relies on standardised terminology that foregrounds technical precision but shows limited engagement with linguistic variation or epistemic context. Researchers in the Russian context operate within a structurally limited discursive environment that suppresses divergence. In both

cases, the capacity to challenge dominant framings or introduce new conceptual tools is reduced. This chapter shows that RS research is not only a technical field but a site of epistemic debate, where language determines who is able to speak, what can be said, and what kinds of agency are permitted.

Chapter 6 traces the conceptual foundations of RS design by analysing how engineering, cognitive science, and behavioural modelling shape the optimisation logic embedded in current systems. The chapter shows how RS development draws on principles of system coherence, predictive efficiency, and user abstraction, inherited from early approaches to machine learning and decision automation. Engineering provides the foundational logic, prioritising performance consistency, system feedback, and reliability. The structuring of RS around algorithmic reward loops, including loss functions and training datasets, reflects assumptions about observable behaviour as a basis for prediction. These assumptions are shaped by disciplinary traditions in which user interaction is formalised as a measurable signal, often detached from social or linguistic context. This paradigm frames optimisation as a technical problem: reducing error, increasing retention, and amplifying engagement. It privileges system performance over communicative diversity. Chapter 6 shows that RS operationalise civic participation through quantifiable proxies, such as click-through rates and watch time, that mask the structural limitations of such metrics in capturing deliberation, dissent, or pluralism. Content that aligns with known user preferences is promoted, while content that is ambiguous, low-frequency, or less reactive is deprioritised. These constraints reflect inherited design priorities that reward behavioural conformity and penalise discursive complexity.

The chapter argues that RS design does not anticipate epistemic agency. It anticipates behavioural regularity. Users are positioned not as contributors to knowledge, but as optimisation targets. This logic narrows the expressive range of RS environments and shapes the linguistic terrain on which civic discourse unfolds. The analysis does not claim intentional exclusion, but shows instead how inherited design parameters constrain what can circulate, be engaged with, or be recognised as meaningful input. RS architectures stabilise certain forms of knowledge expression while filtering others out, not through censorship, but through optimisation procedures aligned with specific system goals. These epistemic effects emerge not only from what RS include, but also from what they structurally omit. By examining how legacy assumptions persist through current RS design, the chapter reframes the epistemic agency gap as an epistemological consequence of optimisation culture. RS do not only shape what is seen. They shape the conditions under which participation occurs. This framing connects system logic to the erosion of epistemic agency in democracy.

This thesis investigates RS as infrastructural systems operating within distinct geopolitical and linguistic contexts: the EU, USA, and Russia, using English and Russian as the primary languages of analysis. These cases were selected to expose structural asymmetries in platform governance, algorithmic design, and research discourse. They do not represent the global diversity of RS deployment. The platforms studied, YouTube and Dzen, serve as examples of how the same underlying technology is implemented differently in specific contexts. This thesis does not address user resistance, informal civic practices, or longitudinal interpretation of RS outputs. It does not capture how users

interpret, ignore, or subvert recommendations in real time. This omission is deliberate, because the thesis does not aim to represent user experience but to analyse how infrastructural conditions shape what forms of experience are made possible. Sock-puppet accounts were not intended to model full human agency. They were designed to isolate recommendation trajectories by holding variables such as language, gender, and search behaviour constant. This made it possible to detect structural asymmetries in visibility, content prioritisation, and discursive range across platforms and geopolitical contexts. The research also does not include ethnographic fieldwork or interviews that would illustrate how users understand or contest algorithmic outputs. Instead, it focuses on system behaviour under controlled input conditions to reveal how RS filter content before any user interpretation takes place. This approach highlights the preconditions of epistemic access. It shows how design choices, ranking logic, and metadata structures determine what becomes available for engagement. Although the study cannot account for all user contexts, it does not seek to generalise. It was designed to examine how RS shape the conditions within which civic participation is enabled or constrained. The methodology combines behavioural simulation, corpus analysis, and regulatory comparison to trace how epistemic agency is structured by platform logic, linguistic configuration, and institutional framing.

RS are not passive intermediaries. They shape what can be known, said, and heard by configuring the flow of information in digital public spheres. As platform algorithms decide which content surfaces, they condition how knowledge power is exercised, increasingly mediating civic discourse through

opaque systems tuned for optimisation. RS structure public discourse by ordering not only content but also the frames, terms, and speakers that define meaning. The sequence in which ideas appear, the repetition of certain framings, and the exclusion of others all consolidate into a logic of visibility, which is neither neutral nor evenly distributed, that defines the epistemic environment. Language plays a central role in shaping these hierarchies. The choice of linguistic input determines not just what is shown, but how civic speech enters the algorithmic field in the first place. When language does not align with dominant training data or metadata conventions, it becomes harder for voices to surface, ideas to gain traction, or publics to appear visible to the system. Engagement metrics exacerbate this condition. Retention, click-through rate, and watch time act as proxies for relevance, sidelining discursive plurality in favour of behavioural predictability. Content that affirms prior assumptions or triggers rapid interaction is amplified, reflecting the optimisation of fast, low-effort responses over slower, more deliberative reasoning described in behavioural economics. Content that introduces complexity, dissent, or slower forms of reasoning is filtered out. As a result, RS systems reward what sustains attention rather than what enables deliberation. This design choice turns civic discourse into a by-product of optimisation, rather than a goal. These problems are not simply functional errors. They are the outcome of a recursive pipeline in which bias is inherited, embedded, and scaled. Developers are trained in environments that treat RS as purely technical systems rather than civic infrastructures. They learn from academic literature shaped by their own epistemic limitations, often English-only, normatively narrow, and ethically abstract. These framings influence how problems are defined, what metrics matter, and which system behaviours are

seen as legitimate. When developers write code, they carry these tacit assumptions into design decisions. These decisions are then hardened into software, baked into interfaces, and deployed at scale. Cultural assumptions also play a role. Developers bring with them localised ideas about what counts as reliable, useful, or acceptable content. These ideas are not universal, but they are often treated as if they were. The result is code that encodes particular perspectives as defaults, making other epistemic approaches harder to recognise. Optimisation does not correct this imbalance. It amplifies it. Once in production, RS systems are steered by economic imperatives and regulatory pressures, with market incentives rewarding systems that maximise retention, segmentation, and monetisation. Regulation, where it exists, focuses on abstract harms such as discrimination or transparency, not on the deeper issue of who gets to speak and be heard. Together, these forces lock in RS designs whose outputs are required to serve commercial or state priorities, rather than democratic needs. Empirical patterns observed across YouTube and Dzen confirm that RS do not operate neutrally across contexts. Lexical variation, sentiment patterns, and gender segmentation align with platform priorities and geopolitical positioning. Across both systems, content visibility is shaped less by civic relevance than by behavioural performance, reinforcing dominant narratives and sidelining linguistic and ideological diversity. Institutional mechanisms for governing RS focus on measurable risks while overlooking epistemic impact. Legal and technical frameworks prioritise compliance over discourse, allowing systems to meet transparency or bias standards while excluding alternative perspectives. The result is regulation that addresses technical error but leaves structural exclusions intact. Academic discourse reinforces this gap by promoting

ambiguous terms such as trustworthy AI or human-in-the-loop, which lack clear definitions but signal ethical intent. These terms enable compliance without structural change, deflecting critique and shielding RS from scrutiny. As a result, both research and regulation fail to address how these systems shape civic life. This is not merely a coordination failure. It is a structural loop. Developers learn from literature that avoids questions of epistemology and the wider social contexts in which RS operate. They build systems in line with optimisation goals. These systems reflect the cultural and normative biases of their creators. Once deployed, they are shaped by market forces and state policy. In the end, RS increasingly shape public knowledge and civic participation, while being evaluated primarily through frameworks oriented toward technical legitimacy. The effect is that RS now exert significant control over epistemic agency without any institutional framework capable of naming or redressing that power.

Recent debates in trust and safety highlight similar concerns. Industry actors now acknowledge that engagement metrics such as time spent and click-through rate are inadequate proxies for civic value. Frameworks such as the DTSP Safe Framework (ISO/IEC 25389) and regulatory initiatives in the EU, UK and Australia signal attempts to impose accountability, yet these measures remain largely procedural (Slachmuis, 2025). They mandate transparency and reporting rather than requiring structural redesign of ranking systems. Without public data access, asymmetries between platforms, regulators and researchers persist, limiting the scope of civic oversight and reinforcing the epistemic constraints already observed in RS.

These dynamics raise the question of how institutional projects are addressing them. The discussion that follows does not evaluate these Horizon Europe projects as completed systems or compare their relative success. Instead, they are treated as illustrative institutional responses to the challenges identified in this thesis, and as reference points for how future projects might be positioned to address epistemic agency at the level of system design rather than at the level of interface or procedure. Among the Horizon Europe initiatives, five are presented here as examples of ongoing institutional efforts to integrate RS governance with broader democratic, environmental and civic objectives: AI4Gov [ai4gov-project.eu], ITHACA [ithaca-project.eu], NeuroClima [neuroclima.eu], ORBIS [orbis-project.eu], and KT4D [kt4democracy.eu]. These projects offer contemporary responses to the ethical, civic and epistemic challenges posed by AI systems in democratic contexts. Emerging from overlapping calls such as HORIZON-CL4-2021-HUMAN01 (EC, 2021), they focus on infrastructures of trust, citizen participation and algorithmic transparency. While each project approaches these priorities differently, all appear grounded in semantic rather than generative architectures. Based on the available documentation, there is little evidence that these systems support dynamic classification from civic input or integrate fine-tuned language models to accommodate discursive plurality. This limits adaptivity and reduces the capacity of publics to shape the epistemic logics embedded in these infrastructures.

AI4Gov (grant agreement 101094905) explores how civic trust in AI-enabled public services can be supported through awareness, transparency and participatory tooling. Coordinated by MAGGIOLI S.p.A., the project develops co-

created digital tools intended to support inclusive and accessible public sector decision-making. It focuses on increasing public understanding of AI technologies through multi-format engagement, including training workshops, awareness-raising events, massive open online courses (MOOCs), and visualisation platforms (AI4Gov, 2024: D5.1; D5.3). The project prioritises explainability through interface-level transparency. Visual tools, such as the Policy Recommendation Toolkit and Bias Detector Catalogue, help users interpret how AI supports policy processes, but deliverables describe these as generating interface-level outputs without user access to underlying classification logic, data schemas, or inference rules (AI4Gov, 2024: D4.3). While the platform supports multilingual design and GDPR compliance, participation features focus on information delivery and feedback collection, without mechanisms to modify embedded semantic structures or ontologies (AI4Gov, 2024: D3.3). Participation is organised around knowledge uptake and interface interaction, not epistemic contribution. Civic engagement is framed as a condition for trust through consultations, workshops, and feedback forms (AI4Gov, 2024: D5.1; D5.2), yet institutional actors retain control over the knowledge architectures by restricting user influence to presentation layers and explanatory outputs rather than core model or data structures (AI4Gov, 2024: D3.3). Public interaction remains limited to surface-level access, with no pathway to reconfigure system logic or introduce alternative epistemic positions. As such, the project expands awareness without redistributing epistemic agency.

ITHACA (grant agreement 101004539) examines how civic participation can inform the democratic governance of AI systems. Coordinated by Università

degli Studi di Modena e Reggio Emilia, the project runs from 2021 to 2024 with pilots in Braşov, Romania, and Martin, Slovakia (ITHACA, 2024: Newsletter #2). Activities in these pilots include the selection of citizen juries, co-design workshops, and participatory algorithmic risk assessments, alongside stakeholder consultations with citizens, municipal staff, and AI practitioners to identify goals, expectations, and barriers to participation (ITHACA, 2025: Newsletter #1; Newsletter #2). These initiatives are multilingual and aim to promote inclusive decision-making and responsible AI deployment in local governance. Project materials describe how citizen input informs discussions of trust, fairness, and accountability, and feeds into policy recommendations (ITHACA, 2025: Newsletter #1). However, deliverables do not indicate that the platform implements retrainable models, grants public access to classification logic, or allows modification of semantic structures. The available documentation describes participation formats that influence process design and interface presentation but provides no evidence of changes to underlying architectures of classification, training, or inference. On this basis, ITHACA can be seen as positioning participation as a mechanism for ethical alignment within existing system boundaries. Civic engagement is procedurally inclusive but does not redistribute control over the infrastructures through which knowledge is produced or circulated.

NeuroClima (grant agreement 101137711) examines how participatory AI tools can support democratic responsiveness in climate governance. Coordinated by Politecnico di Milano, the project runs from 2023 to 2026 and contributes to the EU Adaptation Strategy by fostering behavioural change,

emotional engagement and climate literacy through a socio-technical framework described as a 'nervous system' that connects policymakers, public institutions and citizens (NeuroClima, 2024). Three toolkits called participatory design, creative writing and cinematography, animation and performing arts are deployed in demonstrator pilots using storytelling, co-creation, digital workshops, gamification and performing arts to promote climate dialogue and collaborative learning (NeuroClima, 2024). These activities aim to generate inclusive narratives and shared values around climate adaptation while enhancing trust in institutions (NeuroClima, 2024). Available documentation focuses on participation formats and narrative outcomes but does not extend to technical specifications of how AI operates within the 'nervous system.' This suggests that the projects remain at the level of surface engagement rather than addressing underlying mechanisms such as model retraining, interpretability features, or the integration of civic input into classification and decision-making pipelines. Multilingual and multimodal formats are evident in engagement activities, but there is no evidence of linguistic technologies embedded in the system architecture. On this basis, NeuroClima can be interpreted as foregrounding civic expression and emotional resonance while lacking infrastructures for epistemic intervention or algorithmic reconfiguration.

ORBIS (grant agreement 101094765) seeks to strengthen participation, co-creation, trust, and transparency in deliberative democracy across multiple scales. Coordinated by Politecnico di Milano, the project runs from 2023 to 2026 and develops AI-enabled socio-technical tools for inclusive and scalable democratic engagement (ORBIS, 2023; EU Funding & Tenders Portal, 2025). Its

toolkit combines argument mining, feedback aggregation, explanation generators, and policy recommendation modules that structure debates, synthesise input, and improve interpretability (Cardellino et al., 2024; ORBIS, 2025: Technology). These components are integrated into platforms such as BCause, PolisOrbis, and Democratic Reflection, which support structured dialogue, consensus mapping, and reflective feedback (ORBIS, 2025: Technology; Research). Six pilots at local, national, and European levels apply these tools to topics including security dialogues, youth participation, and EU policy debates (ORBIS, 2025: Pilots). Participation formats are multilingual and multimodal, ranging from digital workshops to hybrid deliberations and narrative engagement (ORBIS, 2025: Research). Conflict-prevention guidelines and argumentation-based analytics are embedded to address risks of polarisation, silence, and lack of common ground (Greco & Jermini-Martinez Soria, 2024; Cardellino et al., 2024). However, available deliverables present mainly surface-level outputs such as summaries, visualisations, and recommendations, without providing public access to classification logics, retrainable models, or semantic structures (ORBIS, 2025: Deliverables). While civic input is aggregated and reflected at the interface, epistemic agency over AI systems remains constrained. ORBIS thus extends inclusivity and visibility in deliberative democracy while retaining institutional and technical control over its infrastructures.

KT4D (grant agreement 101094302), coordinated by Trinity College Dublin, explores how digital transformation can be made accountable to democratic values. The project treats AI governance not as a technical problem

but as a civic and cultural challenge. It centres participatory processes that expose how assumptions are encoded into digital infrastructures. Through the Digital Democracy Lab, KT4D creates spaces where civil society, policymakers and developers engage in collaborative critique of policy narratives, risk framings and governance models (KT4D, 2025). The project promotes multilingual and inclusive participation across societal groups and policy domains. It foregrounds cultural epistemologies and examines how normative assumptions shape system design. Rather than building automated infrastructures, KT4D supports dialogical and narrative approaches to analyse how digital systems mediate civic voice (KT4D, 2025). It experiments with co-created vocabularies and participatory governance methods to rethink classification systems and institutional frameworks. KT4D is distinctive in treating AI as a normative and epistemic structure. It focuses on meaning-making and institutional framing, rather than system performance or interface optimisation. The project positions publics not as users but as co-constructors of the epistemic conditions through which democratic participation becomes possible.

All five projects adopt modular, pilot-based architectures grounded in semantic data models. Public documentation shows no deployment of fine-tuned generative systems. In AI4Gov, ITHACA, and NeuroClima, intervention takes place after the system architecture is defined, with emphasis on outputs, interfaces, and governance processes rather than training data or model logic. This limits the capacity to address structural bias embedded during system development. ORBIS introduces AI-enabled components for structuring debates, aggregating feedback, and generating explanations, but these operate within

predefined models and do not provide mechanisms for public modification of underlying classification logics. Only KT4D includes design features aimed at supporting civic-authored classification structures, although it also stops short of offering an end-to-end pipeline for integrating public input. In all cases, ontologies remain predefined and closed, with minimal scope for public contribution or revision. Legacy semantic tools dominate, and integration of civic-generated taxonomies into classification layers is not prioritised. AI4Gov centres on procedural explainability and institutional usability without infrastructure for epistemic input. ITHACA offers participatory formats but does not embed civic logic in classification. NeuroClima enables expressive engagement but isolates narrative contributions from algorithmic adaptation. KT4D treats classification as a site of democratic agency, with co-authorship of meaning and epistemic structure as explicit design goals. However, the documentation shows no evidence of technical implementation. Civic input is not integrated into pre-processing or model training, where feature selection and loss functions shape classification behaviour. Without this stage-level integration, the approach remains conceptual rather than embedded in system architecture. Across the projects, there is no complete architectural mechanism for integrating public ontologies, folk taxonomies or lived-experience data directly into system logic. Transparency is generally procedural rather than epistemic. KT4D provides the clearest pathways for publics to reframe, dispute or reconfigure governing infrastructures, while the other cases remain bounded by optimisation logics and modular design constraints.

The contrasts drawn across these projects should not be read as a normative ranking or as an assessment of implementation quality. Differences between ORBIS, KT4D, and the other initiatives reflect distinct institutional mandates, stages of development, and research orientations. Open-source availability, interface tooling, or the presence or absence of deployed systems are not treated here as indicators of epistemic agency in themselves. The analysis focuses instead on how civic participation is positioned in relation to classification, optimisation, and meaning-making processes, regardless of delivery model.

The analysis of these Horizon projects reinforces the broader patterns identified across platform outputs, governance frameworks and research discourse in this study. In each case, participation, transparency and inclusivity are presented as priorities, yet underlying classification structures remain closed and resistant to civic intervention. This mirrors the dynamics observed in recommender systems, where core architectures are shaped by inherited developer practices, economic incentives and political alignments that regulators rarely scrutinise and publics have limited means to contest. These conditions restrict civic agency by narrowing the knowledge environment, reinforcing the optimisation logics that already structure public discourse. The dissertation shows how such constraints affect not only what information circulates but also the linguistic diversity through which participation is possible. This erosion of linguistic plurality reduces the capacity to seek, interpret and act on information, deepening the civic empowerment gap. By tracing these mechanisms across empirical platform data, institutional regulation and comparative project analysis,

the study contributes an account of how epistemic agency is shaped in AI-mediated environments and why addressing it requires changes at the level of system design, not only in its procedural interfaces.

This thesis identifies several priorities for further research. First, the sock-puppet methodology could be refined to better capture longitudinal adaptation and interpretive behaviour. Future studies could combine controlled behavioural simulation with complementary qualitative methods, such as diary studies or interviews, to examine how users interpret, negotiate, or resist recommender outputs over time. This would extend the present focus on infrastructural conditions by incorporating user sense-making without abandoning experimental control. Second, regulatory analysis could be expanded to the broader EU digital acquis, particularly the Digital Services Act and subsequent Digital Omnibus initiatives, which post-date the empirical phase of this research. These instruments introduce new transparency, data-access, and systemic risk provisions that may alter opportunities for independent auditing of recommender systems. Future work could examine whether these mechanisms enable substantive scrutiny of optimisation practices or primarily formalise compliance through platform-mediated research access. Third, further analysis of interdisciplinary research discourse is warranted. While this thesis examined technical RS research in English and Russian, future work could extend this approach to legal, policy, and social science literatures to trace how concepts such as accountability, risk, and autonomy travel across disciplines and are transformed in the process. This would allow more precise analysis of boundary-object dynamics beyond technical research communities. Finally, the analytic

chain developed in Chapter 6 offers a framework for future empirical and design-oriented research. Studies could test how interventions at different stages of the chain, such as redefining optimisation objectives, modifying loss functions, or introducing civic input upstream in model development, affect epistemic diversity and agency. Such work would move beyond critique toward experimentally grounded alternatives for recommender system design.

Bibliography

- Abdollahpouri, H., Burke, R. & Mobasher, B. (2019) *Managing Popularity Bias in Recommender Systems with Personalized Re-ranking*. doi:[10.48550/arXiv.1901.07555](https://doi.org/10.48550/arXiv.1901.07555).
- Acemoglu, D. & Robinson, J.A. (2000) Why did the West extend the franchise? Democracy, inequality, and growth in historical perspective. *The quarterly journal of economics*. 115 (4), 1167–1199.
- Adam, A. (2006) *Artificial knowing: Gender and the thinking machine*. Routledge.
<https://www.taylorfrancis.com/books/mono/10.4324/9780203005057/artificial-knowing-alison-adam>.
- Adomavicius, G. & Tuzhilin, A. (2010) Context-aware recommender systems. In: *Recommender systems handbook*. Springer. pp. 217–253.
- Adomavicius, G. & Tuzhilin, A. (2005) Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*. 17 (6), 734–749.
- Aggarwal, C.C. (2016) *Recommender systems*. Springer.
- Agrawal, A., Gans, J. & Goldfarb, A. (2019) *The economics of artificial intelligence: an agenda*. University of Chicago Press.
- Ahearn, L.M. (2001) Language and Agency. *Annual Review of Anthropology*. 30 (1), 109–137. doi:[10.1146/annurev.anthro.30.1.109](https://doi.org/10.1146/annurev.anthro.30.1.109).
- Ahmadian, S., Azarshahi, S. & Paulhus, D.L. (2017) Explaining Donald Trump via communication style: Grandiosity, informality, and dynamism. *Personality and Individual Differences*. 107, 49–53.
- Ahmed, S.T., Basha, S.M., Arumugam, S.R. & Kodabagi, M.M. (2021) *Pattern Recognition: An Introduction*. MileStone Research Publications.
- AI4Gov (2024) *Deliverable D4.3: Policies Visualization Services V1 - AI4Gov*. 12 January 2024. <https://ai4gov-project.eu/2024/01/12/deliverable-d4-3-policies-visualization-services-v1/>.
- Aizenberg, E. & Van Den Hoven, J. (2020) Designing for human rights in AI. *Big Data & Society*. 7 (2), 2053951720949566.
- Akaike, H. (1974) A new look at the statistical model identification. *IEEE transactions on automatic control*. 19 (6), 716–723.

- Akpinar, N.-J. & Fazelpour, S. (2024) *Authenticity and exclusion: social media algorithms and the dynamics of belonging in epistemic communities*. doi:[10.48550/arXiv.2407.08552](https://doi.org/10.48550/arXiv.2407.08552).
- Al Helal, E. & Mokhtar, H. (2018) Towards smart Riyadh: Riyadh Wiki information and complaining system. *International Journal of Managing Information Technology*. 10 (2), 95–106.
- Alex-Assensoh, Y.M. (2005) *Democracy at risk: How political choices undermine citizen participation and what we can do about it*. Brookings Institution Press.
- Alfano, M., Fard, A.E., Carter, J.A., Clutton, P. & Klein, C. (2021) Technologically scaffolded atypical cognition: the case of YouTube's recommender system. *Synthese*. 199 (1–2), 835–858. doi:[10.1007/s11229-020-02724-x](https://doi.org/10.1007/s11229-020-02724-x).
- Alieva, I., Kloo, I. & Carley, K.M. (2024) Analyzing Russia's propaganda tactics on Twitter using mixed methods network analysis and natural language processing: a case study of the 2022 invasion of Ukraine. *EPJ Data Science*. 13 (1), 1–13. doi:[10.1140/epjds/s13688-024-00479-w](https://doi.org/10.1140/epjds/s13688-024-00479-w).
- Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A. & Aljaaf, A.J. (2020) A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In: M.W. Berry, A. Mohamed, & B.W. Yap (eds). *Supervised and Unsupervised Learning for Data Science*. Unsupervised and Semi-Supervised Learning. Cham, Springer International Publishing. pp. 3–21. doi:[10.1007/978-3-030-22475-2_1](https://doi.org/10.1007/978-3-030-22475-2_1).
- Almonte, L., Pérez-Soler, S., Guerra, E., Cantador, I. & De Lara, J. (2021) Automating the synthesis of recommender systems for modelling languages. In: *Proceedings of the 14th ACM SIGPLAN International Conference on Software Language Engineering*. 17 October 2021 Chicago IL USA, ACM. pp. 22–35. doi:[10.1145/3486608.3486905](https://doi.org/10.1145/3486608.3486905).
- Al-Najjar, H.A., Pradhan, B., Beydoun, G., Sarkar, R., Park, H.-J. & Alamri, A. (2023) A novel method using explainable artificial intelligence (XAI)-based Shapley Additive Explanations for spatial landslide prediction using Time-Series SAR dataset. *Gondwana Research*. 123, 107–124.
- Al-Shamri, M.Y.H. (2016) User profiling approaches for demographic recommender systems. *Knowledge-Based Systems*. 100, 175–187.

- Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B. & Zimmermann, T. (2019) Software engineering for machine learning: A case study. In: *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. 2019 IEEE. pp. 291–300. <https://ieeexplore.ieee.org/abstract/document/8804457/>.
- Ananny, M. & Crawford, K. (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*. 20 (3), 973–989. doi:[10.1177/1461444816676645](https://doi.org/10.1177/1461444816676645).
- Anastasiou, L. & Anastasiou, L. (2024) *AI, Big Data and Democracy @ AI Fringe 2024*. doi:[10.5281/zenodo.13929260](https://doi.org/10.5281/zenodo.13929260).
- Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016) *Machine Bias*. 2016. ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [Accessed: 23 July 2025].
- Annoni, A., Benczur, P., Bertoldi, P., Delipetrev, B., De Prato, G., Feijoo, C., Macias, E.F., Gutierrez, E.G., Portela, M.I. & Junklewitz, H. (2018) *Artificial intelligence: A European perspective*. <https://eprints.ugd.edu.mk/28043/1/1.ai-flagship-report-online%20%282%29.pdf>.
- Asmus, R. (2010) *A little war that shook the world: Georgia, Russia, and the future of the West*. St. Martin's Press.
- Atrey, A., Sinha, R., Mitra, S. & Shenoy, P. (2023a) Protecting Proprietary Information in on-Device Machine Learning Models. In: *Proceedings of the 24th International Workshop on Mobile Computing Systems and Applications*. 22 February 2023 Newport Beach California, ACM. pp. 138–138. doi:[10.1145/3572864.3581585](https://doi.org/10.1145/3572864.3581585).
- Atrey, A., Sinha, R., Mitra, S. & Shenoy, P. (2023b) SODA: Protecting Proprietary Information in On-Device Machine Learning Models. In: *Proceedings of the Eighth ACM/IEEE Symposium on Edge Computing*. 6 December 2023 Wilmington DE USA, ACM. pp. 121–132. doi:[10.1145/3583740.3626617](https://doi.org/10.1145/3583740.3626617).
- Augsdorfer, P. (2005) Bootlegging and path dependency. *research policy*. 34 (1), 1–11.
- Axtell, G. (1997) Recent work on virtue epistemology. *American Philosophical Quarterly*. 34 (1), 1–26.

- Azzolina, S., Razza, M., Sartiano, K. & Weitschek, E. (2021) Price Discrimination in the Online Airline Market: An Empirical Study. *Journal of Theoretical and Applied Electronic Commerce Research*. 16 (6), 2282–2303. doi:[10.3390/jtaer16060126](https://doi.org/10.3390/jtaer16060126).
- Baharmand, B. (2023) *Chaos or Clarity: A Corpus-Assisted Analysis of Top Comments on Jordan Peterson's Controversial YouTube Videos*. <https://www.diva-portal.org/smash/record.jsf?pid=diva2:1792294>.
- Bakshy, E., Messing, S. & Adamic, L.A. (2015) Exposure to ideologically diverse news and opinion on Facebook. *Science*. 348 (6239), 1130–1132. doi:[10.1126/science.aaa1160](https://doi.org/10.1126/science.aaa1160).
- Balsa, A.I. & McGuire, T.G. (2001) Statistical discrimination in health care. *Journal of health economics*. 20 (6), 881–907.
- Baltrunas, L. & Amatriain, X. (2009) Towards time-dependant recommendation based on implicit feedback. In: *Workshop on context-aware recommender systems (CARS'09)*. 2009 pp. 25–30.
- Baltrunas, L., Ludwig, B., Peer, S. & Ricci, F. (2012) Context relevance assessment and exploitation in mobile recommender systems. *Personal and Ubiquitous Computing*. 16 (5), 507–526. doi:[10.1007/s00779-011-0417-x](https://doi.org/10.1007/s00779-011-0417-x).
- Baltz, A. (2014) *Through a media lens—The Crimean Crisis: A Discourse Analysis of Media Perspectives on the New Crimean Crisis*. <https://www.diva-portal.org/smash/record.jsf?pid=diva2:746908>.
- Balzer, H. (2003) Managed Pluralism: Vladimir Putin's Emerging Regime. *Post-Soviet Affairs*. 19 (3), 189–227. doi:[10.2747/1060-586X.19.3.189](https://doi.org/10.2747/1060-586X.19.3.189).
- Bania, K. & Geradin, D. (2024) The regulation of cloud computing: why the European Union failed to get it right. *Information & Communications Technology Law*. 33 (1), 99–113. doi:[10.1080/13600834.2023.2260687](https://doi.org/10.1080/13600834.2023.2260687).
- Baraldi, A.N. & Enders, C.K. (2010) An introduction to modern missing data analyses. *Journal of school psychology*. 48 (1), 5–37.
- Barberá, P., Jost, J.T., Nagler, J., Tucker, J.A. & Bonneau, R. (2015) Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*. 26 (10), 1531–1542. doi:[10.1177/0956797615594620](https://doi.org/10.1177/0956797615594620).
- Bardzell, J. & Bardzell, S. (2013) What is 'critical' about critical design? In: *Proceedings of the SIGCHI Conference on Human Factors in*

Computing Systems. 27 April 2013 Paris France, ACM. pp. 3297–3306. doi:[10.1145/2470654.2466451](https://doi.org/10.1145/2470654.2466451).

- Barinov, R., Gai, V., Kuznetsov, G. & Golubenko, V. (2023) Automatic Evaluation of Neural Network Training Results. *Computers*. 12 (2), 26.
- Barnes, A.J., Zhang, Y. & Valenzuela, A. (2024) AI and culture: Culturally dependent responses to AI systems. *Current Opinion in Psychology*. 101838.
- Barnes, R. (2024) 'Isolated nations': the roots of the Russia-North Korea alliance. <https://www.rte.ie/brainstorm/2024/0620/1455665-russia-north-korea-alliance-history-isolated-nations/>.
- Barocas, S., Hardt, M. & Narayanan, A. (2023) *Fairness and machine learning: Limitations and opportunities*. MIT press.
- Barocas, S. & Selbst, A.D. (2016) Big data's disparate impact. *Calif. L. Rev.* 104, 671.
- Barraclough, A. (2024) The best fitness trackers in 2025 to help you stay active. *The Telegraph*. 13 November. <https://www.telegraph.co.uk/recommended/leisure/fitness/best-fitness-trackers-watches/>.
- Batini, C. & Scannapieco, M. (2018) *DATA AND INFORMATION QUALITY: Dimensions, Principles and Techniques*. SPRINGER.
- Battaly, H. (2008) Virtue Epistemology. *Philosophy Compass*. 3 (4), 639–663. doi:[10.1111/j.1747-9991.2008.00146.x](https://doi.org/10.1111/j.1747-9991.2008.00146.x).
- Bauer, G.R. & Scheim, A.I. (2013) Sampling bias in transgender studies. *The Lancet infectious diseases*. 13 (10), 832.
- Baviskar, S. & Malone, M.F.T. (2004) What democracy means to citizens—and why it matters. *Revista Europea de Estudios Latinoamericanos y del Caribe/European Review of Latin American and Caribbean Studies*. 3–23.
- Bawden, D. & Robinson, L. (2009) The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*. 35 (2), 180–191. doi:[10.1177/0165551508095781](https://doi.org/10.1177/0165551508095781).
- BBC (2021) *Capitol riots timeline: What happened on 6 January 2021?* 10 February. <https://www.bbc.com/news/world-us-canada-56004916>.
- BBC (2024a) *Google to fix AI picture bot after 'woke' criticism*. 2024. <https://www.bbc.com/news/business-68364690> [Accessed: 7 July 2025].

- BBC (2024b) *How does the Republican Party pick its presidential nominee?* 4 March. <https://www.bbc.com/news/world-us-canada-66472537>.
- BBC (2024c) *Liverpool to pilot super-smart solar bins that sense when full*. 13 August 2024. <https://www.bbc.com/news/articles/c89w7pxx8n7o> [Accessed: 28 April 2025].
- BBC (2018) *YouTube to label government and public-funded clips*. 8 November. <https://www.bbc.com/news/technology-46139189>.
- Becker, G.S. (1976) *The economic approach to human behavior*. University of Chicago press.
- Beel, J., Gipp, B., Langer, S. & Breiting, C. (2016) Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*. 17 (4), 305–338. doi:[10.1007/s00799-015-0156-0](https://doi.org/10.1007/s00799-015-0156-0).
- Beel, J., Langer, S., Genzmehr, M., Gipp, B., Breiting, C. & Nürnberger, A. (2013) Research paper recommender system evaluation: a quantitative literature survey. In: *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*. RepSys '13. 12 October 2013 New York, NY, USA, Association for Computing Machinery. pp. 15–22. doi:[10.1145/2532508.2532512](https://doi.org/10.1145/2532508.2532512).
- Bender, E.M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 3 March 2021 Virtual Event Canada, ACM. pp. 610–623. doi:[10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922).
- Beniger, J. (1986) *The control revolution: Technological and economic origins of the information society*. Harvard university press.
- Benkler, Y., Faris, R. & Roberts, H. (2018) *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.
- Benov, D.M. (2016) The Manhattan Project, the first electronic computer and the Monte Carlo method. *Monte Carlo Methods and Applications*. 22 (1), 73–79. doi:[10.1515/mcma-2016-0102](https://doi.org/10.1515/mcma-2016-0102).
- Ben-Tal, A., El Ghaoui, L. & Nemirovski, A. (2009) *Robust optimization*. Princeton university press.

- Berelson, B. (1952) Democratic theory and public opinion. *Public Opinion Quarterly*. 313–330.
- Bertolini, A.C. & Schiozer, D.J. (2011) Influence of the objective function in the history matching process. *Journal of Petroleum Science and Engineering*. 78 (1), 32–41.
- Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., Kreitmann, P., Bischof, J. & Chi, E.H. (2019) Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 27 January 2019 Honolulu HI USA, ACM. pp. 453–459.
doi:[10.1145/3306618.3314234](https://doi.org/10.1145/3306618.3314234).
- Beutel, A., Covington, P., Jain, S., Xu, C., Li, J., Gatto, V. & Chi, E.H. (2018) Latent Cross: Making Use of Context in Recurrent Recommender Systems. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 2 February 2018 Marina Del Rey CA USA, ACM. pp. 46–54.
doi:[10.1145/3159652.3159727](https://doi.org/10.1145/3159652.3159727).
- Beygelzimer, A., Langford, J. & Zadrozny, B. (2008) Machine learning techniques—reductions between prediction quality metrics. In: *Performance modeling and engineering*. Springer. pp. 3–28.
- Bhuiyan, J. (2021) LAPD ended predictive policing programs amid public outcry. A new effort shares many of their flaws. *The Guardian*. 8 November. <https://www.theguardian.com/us-news/2021/nov/07/lapd-predictive-policing-surveillance-reform>.
- Biden Jr, J.R. & Carpenter, M. (2018) How to stand up to the Kremlin: Defending democracy against its enemies. *Foreign Aff*. 97, 44.
- Biersdorfer, J.D. (2022) Use That Everyday A.I. in Your Pocket. *The New York Times*. 29 June.
<https://www.nytimes.com/2022/06/29/technology/personaltech/use-that-everyday-ai-in-your-pocket.html>.
- Bietti, E. (2021) From ethics washing to ethics bashing: a moral philosophy view on tech ethics. *Journal of Social Computing*. 2 (3), 266–283.
- Bilsland, B. (2024) *King’s Speech: AI regulation welcome but Labour needs to focus on the bigger picture* | RSM UK. 2024.

<https://www.rsmuk.com/news/ai-regulation-welcome-labour-to-focus-on-bigger-picture> [Accessed: 19 July 2025].

- Binnema, H. & Michels, A. (2022) Does democratic innovation reduce bias? The G1000 as a new form of local citizen participation. *International journal of public administration*. 45 (6), 475–485.
- Binns, R. (2018a) Algorithmic Accountability and Public Reason. *Philosophy & Technology*. 31 (4), 543–556. doi:[10.1007/s13347-017-0263-5](https://doi.org/10.1007/s13347-017-0263-5).
- Binns, R. (2018b) Fairness in machine learning: Lessons from political philosophy. In: *Conference on fairness, accountability and transparency*. 2018 PMLR. pp. 149–159. <https://proceedings.mlr.press/v81/binns18a.html>.
- Binns, R. (2022) Human Judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & Governance*. 16 (1), 197–211. doi:[10.1111/rego.12358](https://doi.org/10.1111/rego.12358).
- Birhane, A. (2021) Algorithmic injustice: a relational ethics approach. *Patterns*. 2 (2). [https://www.cell.com/patterns/fulltext/S2666-3899\(21\)00015-5](https://www.cell.com/patterns/fulltext/S2666-3899(21)00015-5).
- Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R. & Bao, M. (2022) The Values Encoded in Machine Learning Research. In: *2022 ACM Conference on Fairness Accountability and Transparency*. 21 June 2022 Seoul Republic of Korea, ACM. pp. 173–184. doi:[10.1145/3531146.3533083](https://doi.org/10.1145/3531146.3533083).
- Black, J. (2008) Forms and paradoxes of principles-based regulation. *Capital Markets Law Journal*. 3 (4), 425–457.
- Bloch, J. (2006) How to design a good API and why it matters. In: *Companion to the 21st ACM SIGPLAN symposium on Object-oriented programming systems, languages, and applications*. 22 October 2006 Portland Oregon USA, ACM. pp. 506–507. doi:[10.1145/1176617.1176622](https://doi.org/10.1145/1176617.1176622).
- Bloomberg (2020) *Microsoft says it won't sell facial recognition software to police without new regulations*. 11 June 2020. Los Angeles Times. <https://www.latimes.com/business/technology/story/2020-06-11/microsoft-will-stop-selling-facial-recognition-software-to-police> [Accessed: 9 August 2025].

- Blum, A.L. & Rivest, R.L. (1992) Training a 3-node neural network is NP-complete. *Neural Networks*. 5 (1), 117–127.
- Bobadilla, J., Gutiérrez, A., Ortega, F. & Zhu, B. (2018) Reliability quality measures for recommender systems. *Information Sciences*. 442–443, 145–157. doi:[10.1016/j.ins.2018.02.030](https://doi.org/10.1016/j.ins.2018.02.030).
- Bobadilla, J., Ortega, F., Hernando, A. & Gutiérrez, A. (2013) Recommender systems survey. *Knowledge-based systems*. 46, 109–132.
- Bode, L. & Vraga, E.K. (2015) In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*. 65 (4), 619–638.
- Boden, M. (2002) A guide to recurrent neural networks and backpropagation. *the Dallas project*. 2 (2), 1–10.
- Boer, N.-I., Berends, H. & Van Baalen, P. (2011) Relational models for knowledge sharing behavior. *European management journal*. 29 (2), 85–97.
- Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A. & Brunskill, E. (2021) On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. <https://arxiv.org/abs/2108.07258>.
- Bond, M.H., Žegarac, V. & Spencer-Oatey, H. (2000) Culture as an explanatory variable: Problems and possibilities. *Culturally speaking: Managing rapport through talk across cultures*. 293–315.
- Bonicalzi, S., De Caro, M. & Giovanola, B. (2023) Artificial Intelligence and Autonomy: On the Ethical Dimension of Recommender Systems. *Topoi*. 42 (3), 819–832. doi:[10.1007/s11245-023-09922-5](https://doi.org/10.1007/s11245-023-09922-5).
- Borah, A., Mihalcea, R. & Pérez-Rosas, V. (2025) *Persuasion at Play: Understanding Misinformation Dynamics in Demographic-Aware Human-LLM Interactions*. doi:[10.48550/arXiv.2503.02038](https://doi.org/10.48550/arXiv.2503.02038).
- Borodina, S., Deakin, S. & Hamilton, J. (2022) The Rule of Law as an Emergent Social Norm: Evidence from Qualitative Research in Russia. *Law and Development Review*. 15 (1), 1–31. doi:[10.1515/ldr-2021-0063](https://doi.org/10.1515/ldr-2021-0063).
- Bourdieu, P. (1983) The field of cultural production, or: The economic world reversed. *Poetics*. 12 (4–5), 311–356.

- Bowker, G. & Star, S.L. (1999) Sorting things out. *Classification and its consequences*. 4.
<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1903164e809a51032399b8b2c8365572469d5897>.
- Boyle, P.P. (1977) Options: A monte carlo approach. *Journal of financial economics*. 4 (3), 323–338.
- Bozdag, E. & Van Den Hoven, J. (2015) Breaking the filter bubble: democracy and design. *Ethics and Information Technology*. 17 (4), 249–265. doi:[10.1007/s10676-015-9380-y](https://doi.org/10.1007/s10676-015-9380-y).
- Braden, K. (2014) Illegal recreational hunting in Russia: the role of social norms and elite violators. *Eurasian Geography and Economics*. 55 (5), 457–490. doi:[10.1080/15387216.2015.1020320](https://doi.org/10.1080/15387216.2015.1020320).
- Brain, M., Tinelli, C., Rümmer, P. & Wahl, T. (2015) An automatable formal semantics for IEEE-754 floating-point arithmetic. In: *2015 IEEE 22nd Symposium on Computer Arithmetic*. 2015 IEEE. pp. 160–167.
<https://ieeexplore.ieee.org/abstract/document/7203811/>.
- Brauner, P., Glawe, F., Liehner, G.L., Vervier, L. & Ziefle, M. (2024) *Misalignments in AI Perception: Quantitative Findings and Visual Mapping of How Experts and the Public Differ in Expectations and Risks, Benefits, and Value Judgments*. doi:[10.48550/arXiv.2412.01459](https://doi.org/10.48550/arXiv.2412.01459).
- Braunhofer, M., Elahi, M. & Ricci, F. (2014) Usability Assessment of a Context-Aware and Personality-Based Mobile Recommender System. In: M. Hepp & Y. Hoffner (eds). *E-Commerce and Web Technologies*. Lecture Notes in Business Information Processing. Cham, Springer International Publishing. pp. 77–88. doi:[10.1007/978-3-319-10491-1_9](https://doi.org/10.1007/978-3-319-10491-1_9).
- Bray, A.J. & Dean, D.S. (2007) Statistics of critical points of Gaussian fields on large-dimensional spaces. *Physical review letters*. 98 (15), 150201.
- Brittain, B. & Brittain, B. (2023) AI companies ask U.S. court to dismiss artists' copyright lawsuit. *Reuters*. 19 April.
<https://www.reuters.com/legal/ai-companies-ask-us-court-dismiss-artists-copyright-lawsuit-2023-04-19/>.
- Brooke Singman, J.G. (2024) *DOJ IG reveals 26 FBI informants were present on Jan. 6*. 12 December 2024. Fox News.
<https://www.foxnews.com/politics/doj-ig-reveals-number-fbi-confidential-sources-ground-during-jan-6> [Accessed: 2 January 2025].

- Brown, R.D. (2021) Property ownership and the legal personhood of artificial intelligence. *Information & Communications Technology Law*. 30 (2), 208–234. doi:[10.1080/13600834.2020.1861714](https://doi.org/10.1080/13600834.2020.1861714).
- Browne, C.B., Powley, E., Whitehouse, D., Lucas, S.M., Cowling, P.I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S. & Colton, S. (2012) A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*. 4 (1), 1–43.
- Brugnoli, E., Cinelli, M., Zollo, F., Quattrociocchi, W. & Scala, A. (2019) Lexical convergence inside and across echo chambers. *arXiv preprint arXiv:1903.11452*.
- Brusylovska, O. & Maksymenko, I. (2023) Analysis of the media discourse on the 2022 war in Ukraine: The case of Russia. *Regional Science Policy & Practice*. 15 (1), 222–236.
- Bu, Z., Xia, Z. & Wang, J. (2013) A sock puppet detection algorithm on virtual spaces. *Knowledge-Based Systems*. 37, 366–377.
- Buneman, P., Khanna, S. & Wang-Chiew, T. (2001) Why and where: A characterization of data provenance. In: *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings* 8. 2001 Springer. pp. 316–330.
- Burgemeestre, B., Hulstijn, J. & Tan, Y.-H. (2009) Rule-based versus principle-based regulatory compliance. In: *Legal Knowledge and Information Systems*. IOS Press. pp. 37–46.
<https://ebooks.iospress.nl/volumearticle/5520>.
- Burk, D.L. (2019) Algorithmic fair use. *U. Chi. L. Rev.* 86, 283.
- Burke, R. (2002) Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*. 12, 331–370.
- Burke, R., Felfernig, A. & Göker, M.H. (2011) Recommender systems: An overview. *Ai Magazine*. 32 (3), 13–18.
- Burrell, J. (2016) How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*. 3 (1), 205395171562251. doi:[10.1177/2053951715622512](https://doi.org/10.1177/2053951715622512).
- Butenko, E.D., Korolev, V.A., Peredereeva, E.V. & Isakhaev, N.R. (2021) Digital Infrastructure Development in Russia, Europe and Asia. In: E.G. Popkova & B.S. Sergi (eds). *Modern Global Economic System: Evolutional Development vs. Revolutionary Leap*. Lecture Notes in

Networks and Systems. Cham, Springer International Publishing. pp. 1493–1502. doi:[10.1007/978-3-030-69415-9_165](https://doi.org/10.1007/978-3-030-69415-9_165).

- Butler, C. (2018) *Is Google Duplex ethical and moral?* 10 May 2018. Medium. <https://towardsdatascience.com/is-google-duplex-ethical-and-moral-f66a23637640> [Accessed: 5 July 2025].
- Cabral, T.S. (2020) A short guide to the legislative procedure in the European Union. *UNIO-EU Law Journal*. 6 (1), 161–180.
- Cabrio, E., Villata, S. & Cardellino, C. (2024) *From Political Debates to Deliberative Democracy: A Roadmap to Assess Semi-Supervised Argument Mining with DISPUTool*. doi:[10.5281/zenodo.13789192](https://doi.org/10.5281/zenodo.13789192).
- Cakmak, M.C., Agarwal, N. & Oni, R. (2024) The bias beneath: analyzing drift in YouTube’s algorithmic recommendations. *Social Network Analysis and Mining*. 14 (1), 171. doi:[10.1007/s13278-024-01343-5](https://doi.org/10.1007/s13278-024-01343-5).
- Cakmak, M.C., Agarwal, N. & Poudel, D. (2025) *Investigating Algorithmic Bias in YouTube Shorts*. doi:[10.48550/arXiv.2507.04605](https://doi.org/10.48550/arXiv.2507.04605).
- Calma, J. (2023) *Microsoft is going nuclear to power its AI ambitions*. 26 September 2023. The Verge. <https://www.theverge.com/2023/9/26/23889956/microsoft-next-generation-nuclear-energy-smr-job-hiring> [Accessed: 11 July 2025].
- Calo, R. (2011) The boundaries of privacy harm. *Ind. LJ*. 86, 1131.
- Canny, J. (2002) Collaborative filtering with privacy. In: *Proceedings 2002 IEEE Symposium on Security and Privacy*. 2002 IEEE. pp. 45–57.
- Carey, A.N. & Wu, X. (2023) The statistical fairness field guide: perspectives from social and formal sciences. *AI and Ethics*. 3 (1), 1–23. doi:[10.1007/s43681-022-00183-3](https://doi.org/10.1007/s43681-022-00183-3).
- Carlini, N., Feldman, V. & Nasr, M. (2022) *No Free Lunch in ‘Privacy for Free: How does Dataset Condensation Help Privacy’*. doi:[10.48550/arXiv.2209.14987](https://doi.org/10.48550/arXiv.2209.14987).
- Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A. & Raffel, C. (2021) *Extracting Training Data from Large Language Models*. In: 2021 pp. 2633–2650. <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>.

- Castells, P., Hurley, N. & Vargas, S. (2022) Novelty and Diversity in Recommender Systems. In: F. Ricci, L. Rokach, & B. Shapira (eds). *Recommender Systems Handbook*. New York, NY, Springer US. pp. 603–646. doi:[10.1007/978-1-0716-2197-4_16](https://doi.org/10.1007/978-1-0716-2197-4_16).
- Castro, C. & Pham, A.K. (2020) Is the attention economy noxious? *Philosophers' Imprint*. 20 (17), 1–13.
- Cath, C. (2018) Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 376 (2133), 20180080. doi:[10.1098/rsta.2018.0080](https://doi.org/10.1098/rsta.2018.0080).
- Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M. & Floridi, L. (2017) Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and Engineering Ethics*. doi:[10.1007/s11948-017-9901-7](https://doi.org/10.1007/s11948-017-9901-7).
- Cawkell, A.E. (1984) Economics of the information society. *Journal of Information Science*. 8 (1), 42–44.
- Cennamo, C., Kretschmer, T., Constantinides, P., Alaimo, C. & Santaló, J. (2023) Digital platforms regulation: An innovation-centric view of the EU's Digital Markets Act. *Journal of European Competition Law & Practice*. 14 (1), 44–51.
- Central Banking staff (2024) *Cyber security and adapting legacy systems are top tech challenges - Central Banking*. 15 February 2024. <https://www.centralbanking.com/benchmarking/fintech/7960829/cyber-security-and-adapting-legacy-systems-are-top-tech-challenges> [Accessed: 9 July 2025].
- Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. & Mukhopadhyay, D. (2018) *Adversarial Attacks and Defences: A Survey*. doi:[10.48550/arXiv.1810.00069](https://doi.org/10.48550/arXiv.1810.00069).
- Chan, W.K.V. (2013) *Theory and applications of Monte Carlo simulations*. BoD–Books on Demand.
- Chander, A. (2021) *Artificial Intelligence and Trade*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3890420.
- Chander, A. (2016) The racist algorithm. *Mich. L. Rev.* 115, 1023.
- Chaslot, G., Bakkes, S., Szita, I. & Spronck, P. (2008) Monte-carlo tree search: A new framework for game ai. In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital*

Entertainment. 2008 pp. 216–217.

<https://ojs.aaai.org/index.php/AIIDE/article/view/18700>.

- Chassang, G. (2017) The impact of the EU general data protection regulation on scientific research. *ecancermedicalscience*. 11, 709.
- Chawla, D.S. (2022) Massive open index of scholarly papers launches. *Nature*. <https://pubmed.ncbi.nlm.nih.gov/35075274/>.
- Chen, J., Dong, H., Wang, X., Feng, F., Wang, M. & He, X. (2020) Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240*.
- Chen, J., Dong, H., Wang, X., Feng, F., Wang, M. & He, X. (2023) Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*. 41 (3), 1–39.
- Chen, J., Wang, X., Feng, F. & He, X. (2021a) Bias issues and solutions in recommender system: Tutorial on the RecSys 2021. In: *Proceedings of the 15th ACM Conference on Recommender Systems*. 2021 pp. 825–827.
- Chen, L., Zaharia, M. & Zou, J. (2023) *How is ChatGPT's behavior changing over time?* (arXiv:2307.09009). doi:[10.48550/arXiv.2307.09009](https://doi.org/10.48550/arXiv.2307.09009).
- Chen, Q., Ilia, P., Polychronakis, M. & Kapravelos, A. (2021b) Cookie Swap Party: Abusing First-Party Cookies for Web Tracking. In: *Proceedings of the Web Conference 2021*. 19 April 2021 Ljubljana Slovenia, ACM. pp. 2117–2129. doi:[10.1145/3442381.3449837](https://doi.org/10.1145/3442381.3449837).
- Chen, Y., Berkhin, P., Anderson, B. & Devanur, N.R. (2011) Real-time bidding algorithms for performance-based display ad allocation. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 21 August 2011 San Diego California USA, ACM. pp. 1307–1315. doi:[10.1145/2020408.2020604](https://doi.org/10.1145/2020408.2020604).
- Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W. & Ispir, M. (2016) Wide & deep learning for recommender systems. In: *Proceedings of the 1st workshop on deep learning for recommender systems*. 2016 pp. 7–10.
- Cheong, B.C. (2024) Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*. 6, 1421273.
- Cherry, C. (1966) *On human communication*.

- Chilton, P. (2004) *Analysing political discourse: Theory and practice*. routledge.
<https://www.taylorfrancis.com/books/mono/10.4324/9780203561218/analysing-political-discourse-paul-chilton>.
- Cho, J., Ahmed, S., Hilbert, M., Liu, B. & Luu, J. (2020) Do Search Algorithms Endanger Democracy? An Experimental Investigation of Algorithm Effects on Political Polarization. *Journal of Broadcasting & Electronic Media*. 64 (2), 150–172.
doi:[10.1080/08838151.2020.1757365](https://doi.org/10.1080/08838151.2020.1757365).
- Chowdhury, M. & Stoica, I. (2012) Coflow: a networking abstraction for cluster applications. In: *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*. HotNets-XI. 29 October 2012 New York, NY, USA, Association for Computing Machinery. pp. 31–36.
doi:[10.1145/2390231.2390237](https://doi.org/10.1145/2390231.2390237).
- Chu, X., Ilyas, I.F., Krishnan, S. & Wang, J. (2016) Data cleaning: Overview and emerging challenges. In: *Proceedings of the 2016 international conference on management of data*. 2016 pp. 2201–2206.
- Citron, D.K. (2007) Technological due process. *Wash. UL Rev.* 85, 1249.
- Claassen, C. (2020) In the mood for democracy? Democratic support as thermostatic opinion. *American Political Science Review*. 114 (1), 36–53.
- Claessen, E. (2020) Reshaping the internet – the impact of the securitisation of internet infrastructure on approaches to internet governance: the case of Russia and the EU. *Journal of Cyber Policy*. 5 (1), 140–157. doi:[10.1080/23738871.2020.1728356](https://doi.org/10.1080/23738871.2020.1728356).
- Clarivate (2025) *Russian Science Citation Index*. 2025.
<https://webofscience.help.clarivate.com/en-us/Content/russian-science/russian-science-citation-index.htm> [Accessed: 26 February 2025].
- Clarke, J.A. (2024) *Tracking Third-Party Cookies-an empirical analysis of the current situation*. PhD Thesis. York St John University.
<http://ray.yorks.ac.uk/id/eprint/10242/>.
- Cleveland, H. (1985) The twilight of hierarchy: Speculations on the global information society. *Public Administration Review*. 185–195.

- Clinciu, M.A. & Hastie, H.F. (2019) A survey of explainable AI terminology. In: *1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence 2019*. 2019 Association for Computational Linguistics. pp. 8–13.
<https://www.research.ed.ac.uk/en/publications/a-survey-of-explainable-ai-terminology>.
- Coats, S. & Laippala, V. (2024) *Linguistics Across Disciplinary Borders: The March of Data*. Bloomsbury Publishing.
- Cockburn, I.M., Henderson, R. & Stern, S. (2018) The impact of artificial intelligence on innovation: An exploratory analysis. In: *The economics of artificial intelligence: An agenda*. University of Chicago Press. pp. 115–146.
- Coeckelbergh, M. (2020) *AI ethics*. Mit Press.
- Coeckelbergh, M. (2022) Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence. *AI and Ethics*. 1–10.
- Cohen, A. (2009) A value based perspective on commitment in the workplace: An examination of Schwartz’s basic human values theory among bank employees in Israel. *International Journal of Intercultural Relations*. 33 (4), 332–345.
- Cohen, J.E. (2019) *Between truth and power*. Oxford University Press.
- Coleman, J.S. (1988) Social capital in the creation of human capital. *American journal of sociology*. 94, S95–S120.
- Conway, L.G., Gornick, L.J., Houck, S.C., Anderson, C., Stockert, J., Sessoms, D. & McCue, K. (2016) Are Conservatives Really More Simple-Minded than Liberals? The Domain Specificity of Complex Thinking. *Political Psychology*. 37 (6), 777–798.
doi:[10.1111/pops.12304](https://doi.org/10.1111/pops.12304).
- Cornock, M. (2011) Legal definitions of responsibility, accountability and liability: Marc Cornock clarifies the use of terms that are sometimes used interchangeably but have distinct ramifications in law. *Nursing Children and Young People*. 23 (3), 25–26.
doi:[10.7748/ncyp2011.04.23.3.25.c8417](https://doi.org/10.7748/ncyp2011.04.23.3.25.c8417).
- Cottiero, C., Kucharski, K., Olimpieva, E. & Orttung, R.W. (2015) War of words: the impact of Russian state television on the Russian Internet. *Nationalities Papers*. 43 (4), 533–555.

- Covington, P., Adams, J. & Sargin, E. (2016) Deep Neural Networks for YouTube Recommendations. In: *Proceedings of the 10th ACM Conference on Recommender Systems*. 7 September 2016 Boston Massachusetts USA, ACM. pp. 191–198.
doi:[10.1145/2959100.2959190](https://doi.org/10.1145/2959100.2959190).
- Cowling, P.I., Powley, E.J. & Whitehouse, D. (2012) Information set monte carlo tree search. *IEEE Transactions on Computational Intelligence and AI in Games*. 4 (2), 120–143.
- Craglia, M. (2018) *Artificial intelligence: A European perspective*. Publications Office of the European Union.
- Crandell, C.C. & Smaldino, J.J. (2000) Classroom acoustics for children with normal hearing and with hearing impairment. *Language, speech, and hearing services in schools*. 31 (4), 362–370.
- Crawford, K. & Calo, R. (2016) There is a blind spot in AI research. *Nature*. 538 (7625), 311–313.
- Creswell, J.W. & Clark, V.L.P. (2017) *Designing and conducting mixed methods research*. Sage publications.
- Crockett, K., Goltz, S. & Garratt, M. (2018) GDPR impact on computational intelligence research. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. 2018 IEEE. pp. 1–7.
https://ieeexplore.ieee.org/abstract/document/8489614/?casa_token=U9VQ6DjtIIAAAAA:0BfcwelegWp7o6bjZsdhHai9yBOilbSUMVE2mZvtSf0bHZ8MPc3zwalCGrXqfKXqeGFv-6i2IWc.
- Csiszár, I. (2002) The method of types [information theory]. *IEEE Transactions on Information Theory*. 44 (6), 2505–2523.
- Cui, G., Wong, M.L. & Wan, X. (2012) Cost-sensitive learning via priority sampling to improve the return on marketing and CRM investment. *Journal of Management Information Systems*. 29 (1), 341–374.
- Cunningham, W. (1993) The WyCash portfolio management system. *ACM SIGPLAN OOPS Messenger*. 4 (2), 29–30.
doi:[10.1145/157710.157715](https://doi.org/10.1145/157710.157715).
- Curmei, M., Dean, S. & Recht, B. (2021) Quantifying availability and discovery in recommender systems via stochastic reachability. In: *International Conference on Machine Learning*. 2021 PMLR. pp. 2265–2275. <https://proceedings.mlr.press/v139/curmei21a.html>.

- Cusumano, M.A., Mylonadis, Y. & Rosenbloom, R.S. (1992) Strategic maneuvering and mass-market dynamics: The triumph of VHS over Beta. *Business history review*. 66 (1), 51–94.
- Danaher, J. (2016) The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*. 29 (3), 245–268. doi:[10.1007/s13347-015-0211-1](https://doi.org/10.1007/s13347-015-0211-1).
- Danao, M. (2023) 6 Common VPN Protocols Explained. 11 October 2023. Forbes Advisor. <https://www.forbes.com/advisor/business/software/vpn-protocols/> [Accessed: 19 July 2025].
- Daneshjou, R., Smith, M.P., Sun, M.D., Rotemberg, V. & Zou, J. (2021) Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA dermatology*. 157 (11), 1362–1369.
- Daucé, F. & Loveluck, B. (2021) Codes of conduct for algorithmic news recommendation: The Yandex.News controversy in Russia. *First Monday*. doi:[10.5210/fm.v26i5.11708](https://doi.org/10.5210/fm.v26i5.11708).
- Dauphin, Y.N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S. & Bengio, Y. (2014) Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*. 27.
- Dave, K. & Varma, V. (2014) Computational advertising: Techniques for targeting relevant ads. *Foundations and Trends® in Information Retrieval*. 8 (4–5), 263–418.
- Davenport, C. & Armstrong, D.A. (2004) Democracy and the violation of human rights: A statistical analysis from 1976 to 1996. *American Journal of Political Science*. 48 (3), 538–554.
- David, P.A. (1985) Clio and the Economics of QWERTY. *The American economic review*. 75 (2), 332–337.
- Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B. & Sampath, D. (2010) The YouTube video recommendation system. In: *Proceedings of the fourth ACM conference on Recommender systems*. 26 September 2010 Barcelona Spain, ACM. pp. 293–296. doi:[10.1145/1864708.1864770](https://doi.org/10.1145/1864708.1864770).

- De Gregorio, G. (2022) *Digital constitutionalism in Europe: Reframing rights and powers in the algorithmic society*. Cambridge University Press.
- De Gregorio, G. (2021) The rise of digital constitutionalism in the European Union. *International Journal of Constitutional Law*. 19 (1), 41–70.
- DeBray, E.H. (2006) *Politics, ideology, and education*. Teachers College Press.
- Dees, M. (2024) Spotify is restricting developers' API access. 28 November 2024. Techzine Global.
<https://www.techzine.eu/news/applications/126612/spotify-is-restricting-developers-api-access/>.
- Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J. & Jurafsky, D. (2019) *Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings*.
doi:10.48550/arXiv.1904.01596.
- Deng, Y. (2022) Recommender systems based on graph embedding techniques: A review. *IEEE Access*. 10, 51587–51633.
- Denney, E., Pai, G. & Whiteside, I. (2019) The role of safety architectures in aviation safety cases. *Reliability Engineering & System Safety*. 191, 106502.
- Desrosiers, C. & Karypis, G. (2010) A comprehensive survey of neighborhood-based recommendation methods. *Recommender systems handbook*. 107–144.
- Devlin, A.M. (2016) *Positioning The Political Self And Other: Lexical And Syntactic Devices Employed To Position The Other In Russian Online Political Forums*.
https://www.ijors.net/issue5_1_2016/articles/devlin.html.
- Devooght, R. & Bersini, H. (2017) Long and Short-Term Recommendations with Recurrent Neural Networks. In: *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. 9 July 2017 Bratislava Slovakia, ACM. pp. 13–21.
doi:10.1145/3079628.3079670.
- Diamond, L. (1999) *Developing democracy: Toward consolidation*. JHU Press.

- Dickson, D. (1974) *Alternative technology and the politics of technical change*.
- Dignum, V. (2019) *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Artificial Intelligence: Foundations, Theory, and Algorithms. Cham, Springer International Publishing. doi:[10.1007/978-3-030-30371-6](https://doi.org/10.1007/978-3-030-30371-6).
- Dingsøyr, T., Nerur, S., Balijepally, V. & Moe, N.B. (2012) A decade of agile methodologies: Towards explaining agile software development. *Journal of systems and software*. 85 (6) pp.1213–1221. <https://www.sciencedirect.com/science/article/pii/S0164121212000532>.
- Dixon, R. (2024) Under Putin, a militarized new Russia rises to challenge U.S. and the West. *Washington Post*. 6 May. <https://www.washingtonpost.com/world/interactive/2024/putin-values-russian-society-conservatism/>.
- Djeflal, C., Hitrova, C. & Magrani, E. (2021) Recommender Systems and Autonomy: A Role for Regulation of Design, Rights, and Transparency. *Indian JL & Tech*. 17, 1.
- Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., Weller, A. & Wood, A. (2019) *Accountability of AI Under the Law: The Role of Explanation*. doi:[10.48550/arXiv.1711.01134](https://doi.org/10.48550/arXiv.1711.01134).
- Dowty, D. (1991) Thematic proto-roles and argument selection. *language*. 67 (3), 547–619.
- Du, W., Xu, D., Wu, X. & Tong, H. (2021) Fairness-aware agnostic federated learning. In: *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. 2021 SIAM. pp. 181–189.
- Duéñez-Guzmán, E.A., McKee, K.R., Mao, Y., Coppin, B., Chiappa, S., Vezhnevets, A.S., Bakker, M.A., Bachrach, Y., Sadedin, S. & Isaac, W. (2021) Statistical discrimination in learning agents. *arXiv preprint arXiv:2110.11404*.
- Duff, A.S. (2013) *Information society studies*. Routledge.
- Duranti, A. (2005) *A Companion to Linguistic Anthropology*. 1st edition. Wiley. doi:[10.1002/9780470996522](https://doi.org/10.1002/9780470996522).
- Dzmitryieva, A. (2024) *Russian Legal Culture in the Courts and the Judiciary: a Socio-Legal Analysis*. PhD Thesis. Itä-Suomen yliopisto.

https://erepo.uef.fi/bitstream/handle/123456789/31286/urn_isbn_978-952-61-5102-1.pdf?sequence=1.

- Ebers, M., Hoch, V.R., Rosenkranz, F., Ruschemeier, H. & Steinrötter, B. (2021) The European Commission's proposal for an Artificial Intelligence Act—a critical assessment by members of the Robotics and AI Law Society (RAILS). *J. 4* (4), 589–603.
- Ebert, I., Wildhaber, I. & Adams-Prassl, J. (2021) Big Data in the workplace: Privacy Due Diligence as a human rights-based approach to employee privacy protection. *Big Data & Society*. 8 (1), 20539517211013051. doi:[10.1177/20539517211013051](https://doi.org/10.1177/20539517211013051).
- EC (2021) *EU Funding & Tenders Portal*. 2021. <https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl4-2021-human-01-01> [Accessed: 15 June 2025].
- Effing, R., Hillegersberg, J. van & Huibers, T. (2011) Social media and political participation: are Facebook, Twitter and YouTube democratizing our political systems? In: *International conference on electronic participation*. 2011 Springer. pp. 25–35.
- Eidsheim, N.S. (2008) *Voice as a technology of selfhood: Towards an analysis of racialized timbre and vocal performance*. University of California, San Diego.
- Ekstrand, M.D., Das, A., Burke, R. & Diaz, F. (2022) Fairness in Recommender Systems. In: F. Ricci, L. Rokach, & B. Shapira (eds). *Recommender Systems Handbook*. New York, NY, Springer US. pp. 679–707. doi:[10.1007/978-1-0716-2197-4_18](https://doi.org/10.1007/978-1-0716-2197-4_18).
- Ekstrand, M.D., Riedl, J.T. & Konstan, J.A. (2011) Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction*. 4 (2), 81–173.
- Ekstrand, M.D., Tian, M., Azpiazu, I.M., Ekstrand, J.D., Anuyah, O., McNeill, D. & Pera, M.S. (2018a) All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In: S.A. Friedler & C. Wilson (eds). *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. 2018 Proceedings of Machine Learning Research, PMLR. pp. 172–186. <https://proceedings.mlr.press/v81/ekstrand18b.html>.

- Ekstrand, M.D., Tian, M., Kazi, M.R.I., Mehrpouyan, H. & Kluver, D. (2018b) Exploring author gender in book rating and recommendation. In: *Proceedings of the 12th ACM conference on recommender systems*. 2018 pp. 242–250.
- Elger, T. (2024) Braverman, capital accumulation and deskilling 1. In: *The Degradation of Work?*. Routledge. pp. 23–53.
<https://www.taylorfrancis.com/chapters/edit/10.4324/9781003511199-2/braverman-capital-accumulation-deskilling-1-tony-elger>.
- Elger, T. (1979) Valorisation and ‘Deskilling’: A Critique of Braverman. *Capital & Class*. 3 (1), 58–99. doi:[10.1177/030981687900700104](https://doi.org/10.1177/030981687900700104).
- Elgin, C.Z. (2013) Epistemic agency. *Theory and Research in Education*. 11 (2), 135–152. doi:[10.1177/1477878513485173](https://doi.org/10.1177/1477878513485173).
- Elias, M., Tavakoli, M., Lohmann, S., Kismihok, G. & Auer, S. (2020) An OER Recommender System Supporting Accessibility Requirements. In: *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 26 October 2020 Virtual Event Greece, ACM. pp. 1–4. doi:[10.1145/3373625.3418021](https://doi.org/10.1145/3373625.3418021).
- Elkan, C. (2001) The foundations of cost-sensitive learning. In: *International joint conference on artificial intelligence*. 2001 Lawrence Erlbaum Associates Ltd. pp. 973–978.
<http://cseweb.ucsd.edu/~elkan/rescale.pdf>.
- Emde, K., Klimmt, C. & Schluetz, D.M. (2016) Does Storytelling help Adolescents to Process the News?: A comparison of narrative news and the inverted pyramid. *Journalism Studies*. 17 (5), 608–627. doi:[10.1080/1461670X.2015.1006900](https://doi.org/10.1080/1461670X.2015.1006900).
- Engel, P. (2013) IS EPISTEMIC AGENCY POSSIBLE? *Philosophical Issues*. 23 (1), 158–178. doi:[10.1111/phis.12008](https://doi.org/10.1111/phis.12008).
- Englehardt, S., Reisman, D., Eubank, C., Zimmerman, P., Mayer, J., Narayanan, A. & Felten, E.W. (2015) Cookies That Give You Away: The Surveillance Implications of Web Tracking. In: *Proceedings of the 24th International Conference on World Wide Web*. 18 May 2015 Florence Italy, International World Wide Web Conferences Steering Committee. pp. 289–299. doi:[10.1145/2736277.2741679](https://doi.org/10.1145/2736277.2741679).
- Epifanova, A. (2020) *Deciphering Russia’s “Sovereign internet law”: Tightening control and accelerating the Splinternet*.
<https://www.ssoar.info/ssoar/handle/document/66221>.

- Epifanova, A. & Dietrich, P. (2022) *Russia's Quest for Digital Sovereignty: Ambitions, Realities, and Its Place in the World*. DGAP Analysis. Berlin, Forschungsinstitut der Deutschen Gesellschaft für Auswärtige Politik e.V.
- Erbsen, H. & Pöldre, S. (2023) Is all Russian news the same? Framing in Russian news media generated by the Yandex news algorithm for the United States, Estonia, and Russia. *Journalism*. 24 (8), 1789–1816. doi:[10.1177/14648849211069237](https://doi.org/10.1177/14648849211069237).
- Ermoshina, K., Loveluck, B. & Musiani, F. (2022) A market of black boxes: The political economy of Internet surveillance and censorship in Russia. *Journal of Information Technology & Politics*. <https://www.tandfonline.com/doi/abs/10.1080/19331681.2021.1905972>.
- Eskandarian, F., Sonboli, N. & Mobasher, B. (2019) Power of the few: Analyzing the impact of influential users in collaborative recommender systems. In: *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 2019 pp. 225–233.
- EU (2021) *Proposal For A Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>.
- EUR-Lex (2021) *Proposal For A Regulation Of The European Parliament And Of The Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>.
- European Council (2024) *Artificial Intelligence Act: Council calls for promoting safe AI that respects fundamental rights*. 2024. Consilium. <https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/> [Accessed: 8 July 2025].
- Ewold, F. (1991) Insurance and risk. *The Foucault effect: Studies in governmentality*. 197210, 201–202.

- Faddoul, M., Chaslot, G. & Farid, H. (2020) *A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos*. doi:[10.48550/arXiv.2003.03318](https://doi.org/10.48550/arXiv.2003.03318).
- Faisal, N.A. (2023) *Do Banks Price Discriminate Based on Depositors' Location?* doi:[10.2139/ssrn.5038968](https://doi.org/10.2139/ssrn.5038968).
- Fanca, A., Puscasiu, A., Gota, D.-I. & Valean, H. (2020) Recommendation systems with machine learning. In: *2020 21th International Carpathian Control Conference (ICCC)*. 2020 IEEE. pp. 1–6. <https://ieeexplore.ieee.org/abstract/document/9257290/>.
- Färber, M., Lamprecht, D., Krause, J., Aung, L. & Haase, P. (2023) SemOpenAlex: The Scientific Landscape in 26 Billion RDF Triples. In: T.R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, & J. Li (eds). *The Semantic Web – ISWC 2023*. Lecture Notes in Computer Science. Cham, Springer Nature Switzerland. pp. 94–112. doi:[10.1007/978-3-031-47243-5_6](https://doi.org/10.1007/978-3-031-47243-5_6).
- Farer, T.J. (1993) Collectively Defending Democracy in a World of Sovereign States: The Western Hemisphere's Prospect. *Human Rights Quarterly*. 15 (4), 716–750.
- Farzan, R. & Brusilovsky, P. (2011) Encouraging user participation in a course recommender system: An impact on user behavior. *Computers in Human Behavior*. 27 (1), 276–284.
- Fausett, L.V. (2006) *Fundamentals of neural networks: architectures, algorithms and applications*. Pearson Education India.
- Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A. & Kashef, R. (2020) Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *applied sciences*. 10 (21), 7748.
- Fearn, N. (2024) Banks moving into the cloud prompt forecasts of security risk. *Financial Times*. 2 May. <https://www.ft.com/content/2b36a642-bda5-4e43-9747-2175c4d72fd0>.
- Feijóo, C., Kwon, Y., Bauer, J.M., Bohlin, E., Howell, B., Jain, R., Potgieter, P., Vu, K., Whalley, J. & Xia, J. (2020) Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy. *Telecommunications Policy*. 44 (6), 101988.
- Felfernig, A., Jeran, M., Ninaus, G., Reinfrank, F. & Reiterer, S. (2013) Toward the Next Generation of Recommender Systems: Applications and Research Challenges. In: G.A. Tsihrintzis, M. Virvou, & L.C. Jain

- (eds). *Multimedia Services in Intelligent Environments*. Smart Innovation, Systems and Technologies. Heidelberg, Springer International Publishing. pp. 81–98. doi:[10.1007/978-3-319-00372-6_5](https://doi.org/10.1007/978-3-319-00372-6_5).
- Feng, C., Khan, M., Rahman, A.U. & Ahmad, A. (2020) News Recommendation Systems - Accomplishments, Challenges & Future Directions. *IEEE Access*. 8, 16702–16725. doi:[10.1109/ACCESS.2020.2967792](https://doi.org/10.1109/ACCESS.2020.2967792).
 - Feng, S., Park, C.Y., Liu, Y. & Tsvetkov, Y. (2023) From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *arXiv preprint arXiv:2305.08283*.
 - Ferraro, A., Ferreira, G., Diaz, F. & Born, G. (2023) *Commonality in Recommender Systems: Evaluating Recommender Systems to Enhance Cultural Citizenship*. doi:[10.48550/arXiv.2302.11360](https://doi.org/10.48550/arXiv.2302.11360).
 - Ferraro, A., Ferreira, G., Diaz, F. & Born, G. (2024) Measuring Commonality in Recommendation of Cultural Content to Strengthen Cultural Citizenship. *ACM Transactions on Recommender Systems*. 2 (1), 1–32. doi:[10.1145/3643138](https://doi.org/10.1145/3643138).
 - Ferraro, A., Serra, X. & Bauer, C. (2021) Break the loop: Gender imbalance in music recommenders. In: *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 2021 pp. 249–254.
 - Ferrer, X., van Nuenen, T., Such, J.M., Coté, M. & Criado, N. (2021) Bias and discrimination in AI: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*. 40 (2), 72–80.
 - Fink, L., Newman, L. & Haran, U. (2024) Let me decide: Increasing user autonomy increases recommendation acceptance. *Computers in Human Behavior*. 156, 108244.
 - Fiske, A.P. (2004) Four modes of constituting relationships: Consubstantial assimilation; space, magnitude, time, and force; concrete procedures; abstract symbolism. In: *Relational models theory*. Psychology Press. pp. 61–146. <https://www.taylorfrancis.com/chapters/edit/10.4324/9781410611413-5/four-modes-constituting-relationships-alan-page-fiske>.

- Fiske, A.P. (1991) *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing*. Free Press.
- Fiske, S.T., Cuddy, A.J. & Glick, P. (2007) Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*. 11 (2), 77–83.
- Floridi, L. (2018) Soft ethics, the governance of the digital and the General Data Protection Regulation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 376 (2133), 20180081. doi:[10.1098/rsta.2018.0081](https://doi.org/10.1098/rsta.2018.0081).
- Floridi, L. (2021) The European Legislation on AI: a Brief Analysis of its Philosophical Approach. *Philosophy & Technology*. 34 (2), 215–222. doi:[10.1007/s13347-021-00460-9](https://doi.org/10.1007/s13347-021-00460-9).
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. & Vayena, E. (2018) AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*. 28 (4), 689–707. doi:[10.1007/s11023-018-9482-5](https://doi.org/10.1007/s11023-018-9482-5).
- Folorunso, A., Adewumi, T., Adewa, A., Okonkwo, R. & Olawumi, T.N. (2024) Impact of AI on cybersecurity and security compliance. *Global Journal of Engineering and Technology Advances*. 21 (01), 167–184.
- Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D. & Barabási, A.-L. (2018) Science of science. *Science*. 359 (6379), eaao0185. doi:[10.1126/science.aao0185](https://doi.org/10.1126/science.aao0185).
- Foth, M. (2017) The software-sorted city: Big data and algorithms. In: *Digital cities workshop*. 2017 pp. 1–1. <https://eprints.qut.edu.au/107748/>.
- Foucault, M. (2013) *Archaeology of knowledge*. routledge. <https://www.taylorfrancis.com/books/mono/10.4324/9780203604168/archaeology-knowledge-michel-foucault>.
- Foxall, A. (2013) Photographing Vladimir Putin: Masculinity, Nationalism and Visuality in Russian Political Culture. *Geopolitics*. 18 (1), 132–156. doi:[10.1080/14650045.2012.713245](https://doi.org/10.1080/14650045.2012.713245).

- Fraiman, R., Justel, A. & Svarc, M. (2010) Pattern recognition via projection-based kNN rules. *Computational statistics & data analysis*. 54 (5), 1390–1403.
- Frank, J. (2005) *A democracy of distinction: Aristotle and the work of politics*. University of Chicago Press.
- Fraser, H. & y Villarino, J.-M.B. (2023) Acceptable risks in Europe's proposed AI act: reasonableness and other principles for deciding how much risk management is enough. *European Journal of Risk Regulation*. 1–16.
- Fredrikson, M., Jha, S. & Ristenpart, T. (2015) Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. CCS '15. 12 October 2015 New York, NY, USA, Association for Computing Machinery. pp. 1322–1333. doi:[10.1145/2810103.2813677](https://doi.org/10.1145/2810103.2813677).
- Fried, I. (2024) Scoop: Meta won't offer future multimodal AI models in EU. 17 July 2024. Axios. <https://www.axios.com/2024/07/17/meta-future-multimodal-ai-models-eu> [Accessed: 23 July 2025].
- Friedman, B. (1996) Value-sensitive design. *Interactions*. 3 (6), 16–23. doi:[10.1145/242485.242493](https://doi.org/10.1145/242485.242493).
- Friedman, B. (1999) Value-sensitive design: A research agenda for information technology. *Contract No: SBR-9729633*. National Science Foundation, Arlington, VA. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=6d9011f9f3f8f561aa9201f556c2b73a729015c6>.
- Friedman, B., Kahn, P.H., Borning, A. & Hultgren, A. (2013) Value Sensitive Design and Information Systems. In: N. Doorn, D. Schuurbiers, I. Van De Poel, & M.E. Gorman (eds). *Early engagement and new technologies: Opening up the laboratory*. Philosophy of Engineering and Technology. Dordrecht, Springer Netherlands. pp. 55–95. doi:[10.1007/978-94-007-7844-3_4](https://doi.org/10.1007/978-94-007-7844-3_4).
- Friedman, M. (1964) *Capitalism and freedom: With the assistance of Rose D. Friedman*. University of Chicago Press.
- FT Advisor (2024) *Investor interest in ESG drops for third year in a row*. 2024. <https://www.ftadviser.com/investments/2024/10/14/investor-interest-in-esg-drops-for-third-year-in-a-row/> [Accessed: 20 April 2025].

- Fu, M.C. (2016) AlphaGo and Monte Carlo tree search: the simulation optimization perspective. In: *2016 Winter Simulation Conference (WSC)*. 2016 IEEE. pp. 659–670.
https://ieeexplore.ieee.org/abstract/document/7822130/?casa_token=ie hWXOy4XKMAAAAA:TPPRelPoe5t6L4Kkz8hXNfxpPDnpXZtkVdjGJzD 2kdjPqmkE8OK0SChX7aLsTUcJ78zmjR0mv5E.
- Gabriel, I. (2020) Artificial Intelligence, Values, and Alignment. *Minds and Machines*. 30 (3), 411–437. doi:[10.1007/s11023-020-09539-2](https://doi.org/10.1007/s11023-020-09539-2).
- Gal, Y. & Ghahramani, Z. (2016) Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *international conference on machine learning*. 2016 PMLR. pp. 1050–1059.
https://proceedings.mlr.press/v48/gal16.html?trk=public_post_comment-text.
- Galison, P. (2010) Trading with the Enemy. In: M.E. Gorman (ed.). *Trading Zones and International Expertise: Creating New Kinds of Collaboration*. Cambridge, The MIT Press. pp. 25–52.
- Gao, C., Zheng, Y., Li, N., Li, Y., Qin, Y., Piao, J., Quan, Y., Chang, J., Jin, D., He, X. & Li, Y. (2023) A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions. *ACM Transactions on Recommender Systems*. 1 (1), 1–51.
doi:[10.1145/3568022](https://doi.org/10.1145/3568022).
- Garg, P., Villasenor, J. & Foggo, V. (2020) Fairness metrics: A comparative analysis. In: *2020 IEEE international conference on big data (Big Data)*. 2020 IEEE. pp. 3662–3666.
https://ieeexplore.ieee.org/abstract/document/9378025/?casa_token=ov e5ZEBle58AAAAA:fYYdxZt7fbLORAiv10oBz_rHkNe1nxThcwTlwBcPz7 XpmK4jBdrqH8HTDu_pYPpeiZmF3mbO-mY.
- Garrido-Muñoz, I., Montejó-Ráez, A., Martínez-Santiago, F. & Ureña-López, L.A. (2021) A survey on bias in deep NLP. *Applied Sciences*. 11 (7), 3184.
- Garud, R. & Karnoe, P. (2013) *Path dependence and creation*. Psychology Press.
<https://www.taylorfrancis.com/books/mono/10.4324/9781410600370/path-dependence-creation-raghu-garud-peter-karnoe>.
- Gasser, U. & Almeida, V.A. (2017) A layered model for AI governance. *IEEE Internet Computing*. 21 (6), 58–62.

- Gaydareva, I.N., Eshev, M.A. & Markov, P.N. (2020) Internet censorship in the context of legal regulation in Russia. In: *2nd International Scientific and Practical Conference "Modern Management Trends and the Digital Economy: from Regional Development to Global Economic Growth"(MTDE 2020)*. 2020 Atlantis Press. pp. 264–270.
<https://www.atlantis-press.com/proceedings/mtde-20/125939769>.
- Ge, M., Delgado-Battenfeld, C. & Jannach, D. (2010) Beyond accuracy: evaluating recommender systems by coverage and serendipity. In: *Proceedings of the fourth ACM conference on Recommender systems*. 2010 pp. 257–260.
- van Gelder, P., Klaassen, P., Taebi, B., Walhout, B., van Ommen, R., van de Poel, I., Robaey, Z., Asveld, L., Balkenende, R. & Hollmann, F. (2021) Safe-by-design in engineering: An overview and comparative analysis of engineering disciplines. *International Journal of Environmental Research and Public Health*. 18 (12), 6329.
- Gel'man, V. (2015) *Authoritarian Russia: Analyzing post-soviet regime changes*. University of Pittsburgh Press.
- Gera, T. & Singh, J. (2015) A parameterized approach to deal with sock puppets. In: *Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT)*. 2015 IEEE. pp. 1–6.
https://ieeexplore.ieee.org/abstract/document/7060150/?casa_token=nAKR2Wr0c4sAAAAA:6D0rBTa_F-mSXVpDH-5KLj9qSd10sRxDwzUTEbaObDA_6mFMDTIgOo-FileaDS1EO6boTKNXmk.
- Geradin, D. (2018) *What should EU competition policy do to address the concerns raised by the Digital Platforms' market power?*
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3299910.
- Geradin, D., Bania, K. & Karanikioti, T. (2022) The interplay between the Digital Markets Act and the General Data Protection Regulation. *Available at SSRN 4203907*.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4203907.
- Geradin, D., Bania, K., Katsifis, D. & Circiumaru, A. (2022) The regulation of cloud computing: Getting it right. *Available at SSRN 4285731*.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4285731.

- Ghezzi, C. & Guinea, S. (2007) Run-Time Monitoring in Service-Oriented Architectures. In: L. Baresi & E.D. Nitto (eds). *Test and Analysis of Web Services*. Berlin, Heidelberg, Springer Berlin Heidelberg. pp. 237–264. doi:[10.1007/978-3-540-72912-9_9](https://doi.org/10.1007/978-3-540-72912-9_9).
- Gibbons, L. (2025) *State Street and DWS drop ESG from S&P 500 ETFs to align with fund naming rules*. 30 January 2025. ETF Stream. <https://www.etfstream.com/articles/state-street-and-dws-drop-esg-from-s-and-p-500-etfs-to-align-with-fund-naming-rules> [Accessed: 20 April 2025].
- Giessner, S. & Van Quaquebeke, N. (2010) Using a Relational Models Perspective to Understand Normatively Appropriate Conduct in Ethical Leadership. *Journal of Business Ethics*. 95 (S1), 43–55. doi:[10.1007/s10551-011-0790-4](https://doi.org/10.1007/s10551-011-0790-4).
- Giles, K. (2021) Russian information warfare: Construct and purpose. In: *The World Information War*. Routledge. p.
- Gillespie, T. (2018) *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Gillespie, T. (2014) *The relevance of algorithms*. https://direct.mit.edu/books/edited-volume/chapter-pdf/2273545/9780262319461_cai.pdf.
- Gladkova, A. & Ragnedda, M. (2020) Exploring digital inequalities in Russia: an interregional comparative analysis. *Online Information Review*. 44 (4), 767–786.
- Gladkova, A., Vartanova, E. & Ragnedda, M. (2020) Digital divide and digital capital in multiethnic Russian society. *Journal of Multicultural Discourses*. 15 (2), 126–147. doi:[10.1080/17447143.2020.1745212](https://doi.org/10.1080/17447143.2020.1745212).
- Glaeser, E.L., Ponzetto, G.A.M. & Shleifer, A. (2007) Why does democracy need education? *Journal of Economic Growth*. 12 (2), 77–99. doi:[10.1007/s10887-007-9015-1](https://doi.org/10.1007/s10887-007-9015-1).
- Glassman, R.M. (2019) *The future of democracy*. Springer.
- Goddard, M. (2017) The EU General Data Protection Regulation (GDPR): European Regulation that has a Global Impact. *International Journal of Market Research*. 59 (6), 703–705. doi:[10.2501/IJMR-2017-050](https://doi.org/10.2501/IJMR-2017-050).

- Goldberg, D., Nichols, D., Oki, B.M. & Terry, D. (1992) Using collaborative filtering to weave an information tapestry. *Communications of the ACM*. 35 (12), 61–70.
- Gómez, E., Zhang, C.S., Boratto, L., Salamó, M. & Ramos, G. (2022) Enabling cross-continent provider fairness in educational recommender systems. *Future Generation Computer Systems*. 127, 435–447.
- Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep learning*. MIT press.
- Google Ads (2025) *Add age targeting to ad groups - Google Ads Editor Help*. 2025. <https://support.google.com/google-ads/editor/answer/47640?hl=en> [Accessed: 30 August 2025].
- Gorokhovskaia, Y., Shahbaz, A. & Slipowitz, A. (2023) The Freedom House survey for 2022: A turning point for global freedom? *Journal of Democracy*. 34 (2), 105–119.
- Grant, N. (2024) Election Falsehoods Take Off on YouTube as It Looks the Other Way. *The New York Times*. 31 October. <https://www.nytimes.com/2024/10/31/technology/youtube-election-conspiracy-theories-misinformation.html>.
- Grant, N. & Weise, K. (2023) In A.I. Race, Microsoft and Google Choose Speed Over Caution. *The New York Times*. 7 April. <https://www.nytimes.com/2023/04/07/technology/ai-chatbots-google-microsoft.html>.
- Graudina, L.K., Itskovich, V.A. & Katlinskaia, L.P. (2001) Грамматическая правильность русской речи: стилистический словарь вариантов. (No Title). <https://cir.nii.ac.jp/crid/1130282271903491840>.
- Greco, J. & Turri, J. (2017) *Virtue epistemology*. <https://philpapers.org/rec/GREVE-4>.
- Greener, I. (2002) Theorising path-dependency: how does history come to matter in organisations? *Management decision*. 40 (6), 614–619.
- Gritsenko, D., Trochev, A. & Vehkalahti, K. (2025) Public perception of algorithmic policing in a non-democratic context: evidence from Kazakhstan. *Policing and Society*. <https://www.tandfonline.com/doi/abs/10.1080/10439463.2025.2489954>.
- Gross, A., Parker, G. & Fisher, L. (2024) *AI bill and votes at 16 among measures missing from King's Speech*. 2024.

<https://www.ft.com/content/27bb3936-f2e6-4bb3-89e5-5762e4fbf56c>
[Accessed: 19 July 2025].

- Grossmann, M. & Hopkins, D.A. (2015) Ideological republicans and group interest democrats: The asymmetry of American party politics. *Perspectives on Politics*. 13 (1), 119–139.
- Guardian staff (2024) *Super Tuesday 2024: live primary election results | US elections 2024 | The Guardian*. 2024.
<https://www.theguardian.com/us-news/2024/mar/05/super-tuesday-primary-election-live-results-delegates> [Accessed: 1 January 2025].
- Gudmundsson, S. (2013) *General aviation aircraft design: Applied Methods and Procedures*. Butterworth-Heinemann.
- Gunawardana, A. & Shani, G. (2009) A survey of accuracy evaluation metrics of recommendation tasks. *Journal of Machine Learning Research*. 10 (12).
- Gündüz, D., de Kerret, P., Sidiropoulos, N.D., Gesbert, D., Murthy, C.R. & van der Schaar, M. (2019) Machine learning in the air. *IEEE Journal on Selected Areas in Communications*. 37 (10), 2184–2199.
- Gunningham, N. & Sinclair, D. (1999) Integrative regulation: A principle-based approach to environmental policy. *Law & Social Inquiry*. 24 (4), 853–896.
- Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H. & He, Q. (2020) A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*. 34 (8), 3549–3568.
- Gupta, D., Hazarika, B.B. & Berlin, M. (2020) Robust regularized extreme learning machine with asymmetric Huber loss function. *Neural Computing and Applications*. 32 (16), 12971–12998.
- Habermas, J. (1985) *The theory of communicative action: Volume 1: Reason and the rationalization of society*. Beacon press.
- Hacker, P. (2020) *AI Regulation in Europe*. doi:[10.2139/ssrn.3556532](https://doi.org/10.2139/ssrn.3556532).
- Hadar, I., Hasson, T., Ayalon, O., Toch, E., Birnhack, M., Sherman, S. & Balissa, A. (2018) Privacy by designers: software developers' privacy mindset. *Empirical Software Engineering*. 23, 259–289.
- Hadfield-Menell, D. & Hadfield, G.K. (2019) Incomplete Contracting and AI Alignment. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 27 January 2019 Honolulu HI USA, ACM. pp. 417–422. doi:[10.1145/3306618.3314250](https://doi.org/10.1145/3306618.3314250).

- Hadhazy, A. (2023) *ChatGPT Out-scores Medical Students on Complex Clinical Care Exam Questions*. 2023. Stanford HAI.
<https://hai.stanford.edu/news/chatgpt-out-scores-medical-students-complex-clinical-care-exam-questions> [Accessed: 21 July 2025].
- Hagemann, R., Huddleston Skees, J. & Thierer, A. (2018) Soft law for hard problems: The governance of emerging technologies in an uncertain future. *Colo. Tech. LJ*. 17, 37.
- Hagendorff, T. (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*. 30 (1), 99–120. doi:[10.1007/s11023-020-09517-8](https://doi.org/10.1007/s11023-020-09517-8).
- Haim, M., Graefe, A. & Brosius, H.-B. (2018) Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digital journalism*. 6 (3), 330–343.
- Hall, J. & Hofer, C.W. (1993) Venture capitalists' decision criteria in new venture evaluation. *Journal of Business Venturing*. 8 (1), 25–42. doi:[10.1016/0883-9026\(93\)90009-T](https://doi.org/10.1016/0883-9026(93)90009-T).
- Han, X., Baldwin, T. & Cohn, T. (2022) *Balancing out Bias: Achieving Fairness Through Balanced Training*. doi:[10.48550/arXiv.2109.08253](https://doi.org/10.48550/arXiv.2109.08253).
- Hanna, A., Denton, E., Smart, A. & Smith-Loud, J. (2020) Towards a critical race methodology in algorithmic fairness. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 27 January 2020 Barcelona Spain, ACM. pp. 501–512. doi:[10.1145/3351095.3372826](https://doi.org/10.1145/3351095.3372826).
- Hannah, M. (2021a) QAnon and the information dark age. *First Monday*. <http://firstmonday.org/ojs/index.php/fm/article/view/10868>.
- Hannah, M.N. (2021b) A Conspiracy of Data: QAnon, Social Media, and Information Visualization. *Social Media + Society*. 7 (3), 20563051211036064. doi:[10.1177/20563051211036064](https://doi.org/10.1177/20563051211036064).
- Hannak, A., Soeller, G., Lazer, D., Mislove, A. & Wilson, C. (2014) Measuring Price Discrimination and Steering on E-commerce Web Sites. In: *Proceedings of the 2014 Conference on Internet Measurement Conference*. IMC '14. 5 November 2014 New York, NY, USA, Association for Computing Machinery. pp. 305–318. doi:[10.1145/2663716.2663744](https://doi.org/10.1145/2663716.2663744).
- Hansen, L.K., Arvidsson, A., Nielsen, F.A., Colleoni, E. & Etter, M. (2011) Good Friends, Bad News - Affect and Virality in Twitter. In: J.J.

- Park, L.T. Yang, & C. Lee (eds). *Future Information Technology. Communications in Computer and Information Science*. Berlin, Heidelberg, Springer Berlin Heidelberg. pp. 34–43. doi:[10.1007/978-3-642-22309-9_5](https://doi.org/10.1007/978-3-642-22309-9_5).
- Hanson, P. & Teague, E. (2005) Big Business and the State in Russia. *Europe-Asia Studies*. 57 (5), 657–680. doi:[10.1080/09668130500126361](https://doi.org/10.1080/09668130500126361).
 - Hantrais, L. (1999) Contextualization in cross-national comparative research. *International Journal of Social Research Methodology*. 2 (2), 93–108. doi:[10.1080/136455799295078](https://doi.org/10.1080/136455799295078).
 - Hao, K. (2019) This is how AI bias really happens—and why it's so hard to fix. *MIT Technology Review*. 4.
 - Haroon, M., Chhabra, A., Liu, X., Mohapatra, P., Shafiq, Z. & Wojcieszak, M. (2022) YouTube, The Great Radicalizer? Auditing and Mitigating Ideological Biases in YouTube Recommendations. *arXiv preprint arXiv:2203.10666*.
 - Haslam, N. (1995) Factor Structure of Social Relationships: An Examination of Relational Models and Resource Exchange Theories. *Journal of Social and Personal Relationships*. 12 (2), 217–227. doi:[10.1177/0265407595122004](https://doi.org/10.1177/0265407595122004).
 - Haslam, N. (2004) *Relational models theory: A contemporary overview*.
 - Hawkins, P. & McMahon, A. (2020) *Supervision in the Helping Professions 5e*.
 - Hayek, F.A. (2013) The use of knowledge in society. In: *Modern understandings of liberty and property*. Routledge. pp. 27–38. <https://api.taylorfrancis.com/content/chapters/edit/download?identifierName=doi&identifierValue=10.4324/9781315053844-2&type=chapterpdf>.
 - He, H. & Garcia, E.A. (2009) Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*. 21 (9), 1263–1284.
 - He, Z., Zhang, T. & Lee, R.B. (2019) Model inversion attacks against collaborative inference. In: *Proceedings of the 35th Annual Computer Security Applications Conference*. 9 December 2019 San Juan Puerto Rico USA, ACM. pp. 148–162. doi:[10.1145/3359789.3359824](https://doi.org/10.1145/3359789.3359824).
 - Hecht-Nielsen, R. (1992) Theory of the backpropagation neural network. In: *Neural networks for perception*. Elsevier. pp. 65–93.

- Heeks, R. (2002) Information Systems and Developing Countries: Failure, Success, and Local Improvisations. *The Information Society*. 18 (2), 101–112. doi:[10.1080/01972240290075039](https://doi.org/10.1080/01972240290075039).
- Heitmann, B. & Hayes, C. (2010) Using linked data to build open, collaborative recommender systems. In: *2010 AAAI Spring Symposium Series*. 2010 p.
- Helberger, N. (2021) On the democratic role of news recommenders. In: *Algorithms, automation, and news*. Routledge. pp. 14–33. <https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9781003099260-2/democratic-role-news-recommenders-natali-helberger>.
- Helberger, N., Karppinen, K. & D'Acunto, L. (2018) Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*. 21 (2), 191–207. doi:[10.1080/1369118X.2016.1271900](https://doi.org/10.1080/1369118X.2016.1271900).
- Helberger, N., Pierson, J. & Poell, T. (2018) Governing online platforms: From contested to cooperative responsibility. *The Information Society*. 34 (1), 1–14. doi:[10.1080/01972243.2017.1391913](https://doi.org/10.1080/01972243.2017.1391913).
- Helm, P., Bella, G., Koch, G. & Giunchiglia, F. (2023) *Diversity and Language Technology: How Techno-Linguistic Bias Can Cause Epistemic Injustice*. doi:[10.48550/arXiv.2307.13714](https://doi.org/10.48550/arXiv.2307.13714).
- Hennig, C. & Kutlukaya, M. (2007) Some thoughts about the design of loss functions. *REVSTAT-Statistical Journal*. 5 (1), 19-39-19–39.
- Herlocker, J.L., Konstan, J.A., Terveen, L.G. & Riedl, J.T. (2004) Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*. 22 (1), 5–53.
- Herman, E.S. & Chomsky, N. (2010) *Manufacturing consent: The political economy of the mass media*. Random House.
- Herrera-Poyatos, A., Ser, J.D., Prado, M.L. de, Wang, F.-Y., Herrera-Viedma, E. & Herrera, F. (2025) *Responsible Artificial Intelligence Systems: A Roadmap to Society's Trust through Trustworthy AI, Auditability, Accountability, and Governance*. doi:[10.48550/arXiv.2503.04739](https://doi.org/10.48550/arXiv.2503.04739).
- Hildebrandt, M. & Tielemans, L. (2013) Data protection by design and technology neutral law. *Computer Law & Security Review*. 29 (5), 509–521.

- Hiller, J. (2023) *WSJ News Exclusive | Microsoft Targets Nuclear to Power AI Operations*. 12 December 2023. WSJ.
<https://www.wsj.com/tech/ai/microsoft-targets-nuclear-to-power-ai-operations-e10ff798> [Accessed: 11 July 2025].
- Hinote, B.P. & Webber, G.R. (2012) Drinking toward Manhood: Masculinity and Alcohol in the Former USSR. *Men and Masculinities*. 15 (3), 292–310. doi:[10.1177/1097184X12448466](https://doi.org/10.1177/1097184X12448466).
- Hinton, G.E. & Salakhutdinov, R.R. (2006) Reducing the dimensionality of data with neural networks. *science*. 313 (5786), 504–507.
- Ho, Y. & Wookey, S. (2019) The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE access*. 8, 4806–4813.
- Hoda, R., Noble, J. & Marshall, S. (2012) Self-organizing roles on agile software development teams. *IEEE Transactions on Software Engineering*. 39 (3), 422–444.
- Hofstadter, D.R. (1980) Reductionism and religion. *Behavioral and Brain Sciences*. 3 (3), 433–434.
- Hofstede, G. (1984) *Culture's consequences: International differences in work-related values*. sage.
- Hofstede, G. (2011) Dimensionalizing cultures: The Hofstede model in context. *Online readings in psychology and culture*. 2 (1), 8.
- Hoo, Z.H., Candlish, J. & Teare, D. (2017) What is an ROC curve? *Emergency Medicine Journal*. 34 (6) pp.357–359.
https://emj.bmj.com/content/34/6/357.abstract?casa_token=36sX9prlU84AAAAA:5kJj2Gg_3h0vltFeJuB1C-K5ep2m2vcwWRwvV-dhYkljeyPq4q_tlUCaz2L2hGy3XB9ihfkeoMcr.
- Hopper, P.J. & Thompson, S.A. (1980) Transitivity in grammar and discourse. *language*. 56 (2), 251–299.
- Horkheimer, M., Adorno, T.W. & Noeri, G. (2002) *Dialectic of enlightenment*. Stanford University Press.
- Hou, J., Kang, J. & Qi, N. (2010) On Vocabulary Size in Bag-of-Visual-Words Representation. In: G. Qiu, K.M. Lam, H. Kiya, X.-Y. Xue, C.-C.J. Kuo, & M.S. Lew (eds). *Advances in Multimedia Information Processing - PCM 2010*. Lecture Notes in Computer Science. Berlin, Heidelberg, Springer Berlin Heidelberg. pp. 414–424. doi:[10.1007/978-3-642-15702-8_38](https://doi.org/10.1007/978-3-642-15702-8_38).

- House, F. (2018) *Freedom in the World 2018-Russia*.
<https://www.ecoi.net/en/document/1428824.html>.
- House, F. (2017) *Freedom of the Press 2017-Russia*.
<https://www.ecoi.net/en/document/1410423.html>.
- House, R.J., Hanges, P.J., Javidan, M., Dorfman, P.W. & Gupta, V. (2004) *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications.
- Howells, L. & Henry, L.A. (2021) Varieties of digital authoritarianism: Analyzing Russia's approach to Internet governance. *Communist and Post-Communist Studies*. 54 (4), 1–27.
- Huang, J., Oosterhuis, H., De Rijke, M. & Van Hoof, H. (2020) Keeping Dataset Biases out of the Simulation: A Debiased Simulator for Reinforcement Learning based Recommender Systems. In: *Fourteenth ACM Conference on Recommender Systems*. 22 September 2020 Virtual Event Brazil, ACM. pp. 190–199. doi:[10.1145/3383313.3412252](https://doi.org/10.1145/3383313.3412252).
- Hughes, A. (2022) AI Models vs. AI Systems: Understanding Units of Performance Assessment. 19 September 2022. Microsoft Research.
<https://www.microsoft.com/en-us/research/blog/ai-models-vs-ai-systems-understanding-units-of-performance-assessment/>.
- Hughes, T.P. (1987) The evolution of large technological systems. *The social construction of technological systems: New directions in the sociology and history of technology*. 82, 51–82.
- Hupont, I. & Fernández, C. (2019) Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In: *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. 2019 IEEE. pp. 1–7.
- IEEE (2019) *Ethically Aligned Design report*.
- IEEE (2017) *Ethically Aligned Design First Edition Glossary*.
- Illich, I. & Allich, A. (1975) *Medical nemesis*. Australian Broadcasting Commission, Science Programmes Unit Sydney.
<https://www.truthvisionary.com/public/gallery/MEDICAL-NEMESIS-Ivan-Illich.pdf>.
- Illich, I. & Lang, A. (1973) *Tools for conviviality*.
[https://pages.sandpoints.org/ubu50/library/Ivan%20Illich/Tools%20for%20Conviviality%20\(272\)/Tools%20for%20Conviviality%20-%20Ivan%20Illich.pdf](https://pages.sandpoints.org/ubu50/library/Ivan%20Illich/Tools%20for%20Conviviality%20(272)/Tools%20for%20Conviviality%20-%20Ivan%20Illich.pdf).

- Incorvaia, A.D., Fealing, K.H. & Schiebinger, L. (2023) Gender and innovation: indicators and measurement gaps. In: *Handbook of Innovation Indicators and Measurement*. Edward Elgar Publishing. pp. 278–296.
<https://www.elgaronline.com/edcollchap/book/9781800883024/book-part-9781800883024-26.xml>.
- Interfax (2024) *Мишустин заявил об отличии ‘мозгов’ нейросетевых моделей в РФ и на Западе*. 2 February 2024. Interfax.ru. <https://www.interfax.ru/digital/944191> [Accessed: 15 July 2025].
- ITHACA (2025) *Results* – *ithaca-project.eu*. 2025. <https://www.ithaca-project.eu/results/>.
- Jabbar, H. & Khan, R.Z. (2015) Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices*. 70 (10.3850), 978–981.
- Jacobson, A.D. (2005) *The first Gospel: An introduction to Q*. Wipf and Stock Publishers.
- Jacoby, S. (2009) *The age of American unreason*. Vintage.
- Jain, R.K., Chiu, D.-M.W. & Hawe, W.R. (1984) A quantitative measure of fairness and discrimination. *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*. 21, 1.
- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) *An introduction to statistical learning*. Springer.
- Jamieson, K.H. & Cappella, J.N. (2008) *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.
- Jang, H. (2024) *An Integrative Framing Study of the Public Discourse around Artificial Intelligence (AI) Ethics*.
<https://cdr.lib.unc.edu/concern/dissertations/v692tk535>.
- Jannach, D. & Adomavicius, G. (2016) Recommendations with a Purpose. In: *Proceedings of the 10th ACM Conference on Recommender Systems*. 7 September 2016 Boston Massachusetts USA, ACM. pp. 7–10. doi:[10.1145/2959100.2959186](https://doi.org/10.1145/2959100.2959186).
- Jasanoff, S. (2005) *Designs on Nature: Science and Democracy in Europe and the United States*. Princeton, Princeton University Press. doi:[10.1515/9781400837311](https://doi.org/10.1515/9781400837311).

- Jelinek, T. (2023) *The Digital Sovereignty Trap: Avoiding the Return of Silos and a Divided World*. Springer Nature.
- Jennings, T. (2021) *The State and Big Business in Russia: Understanding Kremlin–Business Relations in the Early Putin Era*. Routledge.
<https://www.taylorfrancis.com/books/mono/10.4324/9781003251798/state-big-business-russia-tina-jennings>.
- Jermini, C. & Greco, S. (2024) *Integrating conflict prevention tools into deliberative democracy online platforms*.
doi:[10.5281/zenodo.13832640](https://doi.org/10.5281/zenodo.13832640).
- Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., et al. (2024) *AI Alignment: A Comprehensive Survey*. <http://arxiv.org/abs/2310.19852>.
- Jiang, H., Beeferman, D., Brannon, W., Heyward, A. & Roy, D. (2024) Bridging Dictionary: AI-Generated Dictionary of Partisan Language Use. In: *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. 11 November 2024 San Jose Costa Rica, ACM. pp. 79–82.
doi:[10.1145/3678884.3681820](https://doi.org/10.1145/3678884.3681820).
- Jiang, Y., Lin, J., Cukic, B. & Menzies, T. (2009) Variance analysis in software fault prediction models. In: *2009 20th International Symposium on Software Reliability Engineering*. 2009 IEEE. pp. 99–108.
https://ieeexplore.ieee.org/abstract/document/5362090/?casa_token=Qg1XERPUTYAAAAA:JBZiHvRyrnHANsJJvlyB6mVD_xrm3G0SEUuURIZ0LEv16Up0-hikj42FUBpDEVJ0vll_UHu2Vc0.
- Joachims, T. (1998) Text categorization with support vector machines: Learning with many relevant features. In: *European conference on machine learning*. 1998 Springer. pp. 137–142.
- Jobin, A., Ienca, M. & Vayena, E. (2019) The global landscape of AI ethics guidelines. *Nature machine intelligence*. 1 (9), 389–399.
- Johnson, D.G. & Verdicchio, M. (2017) Reframing AI Discourse. *Minds and Machines*. 27 (4), 575–590. doi:[10.1007/s11023-017-9417-6](https://doi.org/10.1007/s11023-017-9417-6).
- Jones, S.M., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R. & Grover, C. (2016) Scholarly context adrift: three out of four URI references lead to changed content. *PloS one*. 11 (12), e0167475.
- Jost, J.T. (2006) The end of the end of ideology. *American psychologist*. 61 (7), 651.

- Justwan, F., Baumgaertner, B., Carlisle, J.E., Clark, A.K. & Clark, M. (2018) Social media echo chambers and satisfaction with democracy among Democrats and Republicans in the aftermath of the 2016 US elections. *Journal of Elections, Public Opinion and Parties*. 28 (4), 424–442. doi:[10.1080/17457289.2018.1434784](https://doi.org/10.1080/17457289.2018.1434784).
- Kahan, D.M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L.L., Braman, D. & Mandel, G. (2012) The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature climate change*. 2 (10), 732–735.
- Kaiser, J. & Rauchfleisch, A. (2020) Birds of a Feather Get Recommended Together: Algorithmic Homophily in YouTube’s Channel Recommendations in the United States and Germany. *Social Media + Society*. 6 (4), 205630512096991. doi:[10.1177/2056305120969914](https://doi.org/10.1177/2056305120969914).
- Kalathil, S. (2020) The Evolution of Authoritarian Digital Influence: Grappling with the New Normal. *PRISM*. 9 (1), 32–51.
- Kaldestad, Ø. (2024) The Norwegian Consumer Council files legal complaint against Meta for numerous violations of the GDPR. 29 February 2024. Forbrukerrådet. <https://www.forbrukerradet.no/side/the-norwegian-consumer-council-files-legal-complaint-against-meta-for-numerous-violations-of-the-gdpr/>.
- Kaminski, M.E. (2021) The right to explanation, explained. In: *Research Handbook on Information Law and Governance*. Edward Elgar Publishing. pp. 278–299. <https://www.elgaronline.com/edcollchap/edcoll/9781788119917/9781788119917.00024.xml>.
- Kamiya, M. & Keate, J. (2024) *AI Watch: Global regulatory tracker - Japan* | White & Case LLP. 1 July 2024. <https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-japan> [Accessed: 19 July 2025].
- Karp, R.M. (1972) Reducibility among combinatorial problems. In: *Complexity of computer computations*. Springer. pp. 85–103.
- Katayev, A., Balciza, C. & Seccombe, D.W. (2010) Establishing Reference Intervals for Clinical Laboratory Test Results: Is There a Better Way? *American Journal of Clinical Pathology*. 133 (2), 180–186. doi:[10.1309/AJCPN5BMTSF1CDYP](https://doi.org/10.1309/AJCPN5BMTSF1CDYP).

- Kay, A. (2005) A Critique of the Use of Path Dependency in Policy Studies. *Public Administration*. 83 (3), 553–571. doi:[10.1111/j.0033-3298.2005.00462.x](https://doi.org/10.1111/j.0033-3298.2005.00462.x).
- Kay, A. (2003) Path dependency and the CAP. *Journal of European Public Policy*. 10 (3), 405–420. doi:[10.1080/1350176032000085379](https://doi.org/10.1080/1350176032000085379).
- Kay, A. (2007) Tesseract: an open-source optical character recognition engine. *Linux Journal*. 2007 (159), 2.
- Kayam, O. (2018) The Readability and Simplicity of Donald Trump's Language. *Political Studies Review*. 16 (1), 73–88. doi:[10.1177/1478929917706844](https://doi.org/10.1177/1478929917706844).
- Kaznatcheev, A. & Kording, K.P. (2022) Nothing makes sense in deep learning, except in the light of evolution. *arXiv preprint arXiv:2205.10320*.
- Ke, T.T. & Sudhir, K. (2023) Privacy Rights and Data Security: GDPR and Personal Data Markets. *Management Science*. 69 (8), 4389–4412. doi:[10.1287/mnsc.2022.4614](https://doi.org/10.1287/mnsc.2022.4614).
- Keith, W.M. (2007) *Democracy as discussion: Civic education and the American forum movement*. Lexington Books.
- Keller, K.L. & Staelin, R. (1987) Effects of Quality and Quantity of Information on Decision Effectiveness. *Journal of Consumer Research*. 14 (2), 200–213. doi:[10.1086/209106](https://doi.org/10.1086/209106).
- Kendall, A. & Gal, Y. (2017) What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*. 30. <https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html>.
- Kerr, J. (2019) *The Russian Model of Digital Control and Its Significance*. pp.62–74. <https://www.jstor.org/stable/resrep19585.14>.
- Khanal, S.S., Prasad, P.W.C., Alsadoon, A. & Maag, A. (2020) A systematic review: machine learning based recommendation systems for e-learning. *Education and Information Technologies*. 25 (4), 2635–2664. doi:[10.1007/s10639-019-10063-9](https://doi.org/10.1007/s10639-019-10063-9).
- Kharitonova, Y., Malik, N.S. & Yang, T. (2023) The Legal Issue Of Deterrence Of Algorithmic Control Of Digital Platforms: The Experience Of China, The European Union, Russia And India. *BRICS Law Journal*. 10 (1), 147–170.

- Kharitonova, Y.S. (2023) Legal Means of Providing the Principle of Transparency of the Artificial Intelligence. *Journal of Digital Technologies and Law*. 1 (2). <https://cyberleninka.ru/article/n/legal-means-of-providing-the-principle-of-transparency-of-the-artificial-intelligence>.
- Khusro, S., Ali, Z. & Ullah, I. (2016) Recommender Systems: Issues, Challenges, and Research Opportunities. In: K.J. Kim & N. Joukov (eds). *Information Science and Applications (ICISA) 2016*. 2016 Singapore, Springer Singapore. pp. 1179–1189.
- Kim, M.P., Ghorbani, A. & Zou, J. (2019) Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 27 January 2019 Honolulu HI USA, ACM. pp. 247–254.
doi:[10.1145/3306618.3314287](https://doi.org/10.1145/3306618.3314287).
- Kinyanjui, N.M., Odonga, T., Cintas, C., Codella, N.C.F., Panda, R., Sattigeri, P. & Varshney, K.R. (2019) *Estimating Skin Tone and Effects on Classification Performance in Dermatology Datasets*.
doi:[10.48550/arXiv.1910.13268](https://doi.org/10.48550/arXiv.1910.13268).
- Klappholz, S. (2024) *Google settles \$5 billion privacy lawsuit ahead of cookie phase-out*. 2 January 2024. ITPro.
<https://www.itpro.com/software/google/google-settles-dollar5-billion-privacy-lawsuit-ahead-of-cookie-phase-out> [Accessed: 12 July 2025].
- Kleijn, W.B., Storus, A., Chinen, M., Denton, T., Lim, F.S., Luebs, A., Skoglund, J. & Yeh, H. (2021) Generative speech coding with predictive variance regularization. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021 IEEE. pp. 6478–6482.
https://ieeexplore.ieee.org/abstract/document/9415120/?casa_token=2R8FpGHnHRQAAAAA:gBSTclBMwDU1vumyecXCQ8wBkraLulcz-gWZIJx2ij-L0EQ8zyQry-JgqWyVtVpj9HXxv3d5Ro.
- Knees, P., Neidhardt, J. & Nalis, I. (2024) Recommender Systems: Techniques, Effects, and Measures Toward Pluralism and Fairness. *Hannes Werthner· Carlo Ghezzi· Jeff Kramer· Julian Nida-Rümelin· Bashar Nuseibeh· Erich Prem*. 417.
- Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H. & Newell, C. (2012) Explaining the user experience of recommender systems.

User Modeling and User-Adapted Interaction. 22 (4–5), 441–504.
doi:[10.1007/s11257-011-9118-4](https://doi.org/10.1007/s11257-011-9118-4).

- Knowles, B. & Richards, J.T. (2021) *The Sanction of Authority: Promoting Public Trust in AI*. doi:[10.48550/arXiv.2102.04221](https://doi.org/10.48550/arXiv.2102.04221).
- Ko, H., Lee, S., Park, Y. & Choi, A. (2022) A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*. 11 (1), 141.
- Koehrsen, W. (2018) Overfitting vs. underfitting: A complete example. *Towards Data Science*. 405.
- Kolmogorov, A. (1956) On the Shannon theory of information transmission in the case of continuous signals. *IRE Transactions on Information Theory*. 2 (4), 102–108.
- Konam, S. (2022) *Where did IBM go wrong with Watson Health?* 2 March 2022. Quartz. <https://qz.com/2129025/where-did-ibm-go-wrong-with-watson-health> [Accessed: 5 July 2025].
- Konstan, J.A. & Riedl, J. (2012) Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*. 22 (1–2), 101–123. doi:[10.1007/s11257-011-9112-x](https://doi.org/10.1007/s11257-011-9112-x).
- Koops, B.-J. (2006) Should ICT regulation be technology-neutral? *Starting points for ict regulation. Deconstructing prevalent policy one-liners, it & law series, Bert-Jaap Koops, Miriam Lips, Corien Prins & Maurice Schellekens, eds.* 9, 77–108.
- Koren, Y. (2009) Collaborative filtering with temporal dynamics. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009 pp. 447–456.
- Korte, R.F. (2003) Biases in decision making and implications for human resource development. *Advances in Developing Human Resources*. 5 (4), 440–457.
- Kostyuk, N. & Zhukov, Y.M. (2019) Invisible Digital Front: Can Cyber Attacks Shape Battlefield Events? *Journal of Conflict Resolution*. 63 (2), 317–347. doi:[10.1177/0022002717737138](https://doi.org/10.1177/0022002717737138).
- Krafft, P.M., Young, M., Katell, M., Huang, K. & Bugingo, G. (2019) *Defining AI in Policy versus Practice*. doi:[10.48550/arXiv.1912.11095](https://doi.org/10.48550/arXiv.1912.11095).
- Kremlin (2018) *Указ Президента Российской Федерации от 06.12.2018 № 703 · Официальное опубликование правовых актов*. 2018.

<http://publication.pravo.gov.ru/Document/View/0001201812070007>

[Accessed: 16 July 2025].

- Kremlin (2024) *Указ Президента РФ от 10.10.2019 N 490 (ред. от 15.02.2024)*. <https://www.kremlin.ru/acts/bank/44731>.
- Krige, J. & Pestre, D. (2013) *Science in the twentieth century*. Routledge.
<https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.4324/9781315079097&type=googlepdf>.
- KT4D (2025) *Publications KT4democracy*. 8 May 2025.
<https://kt4democracy.eu/publications> [Accessed: 14 August 2025].
- Kuanr, M. & Mohapatra, P. (2021) Assessment Methods for Evaluation of Recommender Systems: A Survey. *Foundations of Computing and Decision Sciences*. 46 (4), 393–421. doi:[10.2478/fcds-2021-0023](https://doi.org/10.2478/fcds-2021-0023).
- Kunaver, M. & Požrl, T. (2017) Diversity in recommender systems—A survey. *Knowledge-based systems*. 123, 154–162.
- Kurkchiyan, M. (2009) Russian legal culture: an analysis of adaptive response to an institutional transplant. *Law & Social Inquiry*. 34 (2), 337–364.
- Kuznetsova, E. & Makhortykh, M. (2023) Blame It on the Algorithm? Russian Government-Sponsored Media and Algorithmic Curation of Political Information on Facebook. *International Journal of Communication*. 17 (0), 22.
- LaCroix, T. & Mohseni, A. (2021) The tragedy of the AI commons. *Synthese*. 200 (4), 1–33.
- Lai, S., Potter, Y., Kim, J., Zhuang, R., Song, D. & Evans, J. (2024) *Evolving AI Collectives to Enhance Human Diversity and Enable Self-Regulation*. <http://arxiv.org/abs/2402.12590>.
- Lalla, V., Meyer, S.P., Tejani, Z., Hayes, D.L., Mitrani, A. & Tohme, A. (2023) *Key Provisions and Impacts of Biden’s Executive Order on AI...* 9 November 2023. Fenwick.
<https://www.fenwick.com/insights/publications/key-provisions-and-impacts-of-bidens-executive-order-on-ai-regulation-and-development> [Accessed: 8 July 2025].
- Lathia, N., Hailes, S., Capra, L. & Amatriain, X. (2010) Temporal diversity in recommender systems. In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in*

information retrieval. 19 July 2010 Geneva Switzerland, ACM. pp. 210–217. doi:[10.1145/1835449.1835486](https://doi.org/10.1145/1835449.1835486).

- Lau, P.L. (2019) The extension of legal personhood in artificial intelligence. *Rev. Bioetica & Derecho*. 46, 47.
- Ledwich, M. & Zaitsev, A. (2019) *Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization*. <http://arxiv.org/abs/1912.11211>.
- Lee, K. & Joshi, K. (2020) Understanding the Role of Cultural Context and User Interaction in Artificial Intelligence Based Systems. *Journal of Global Information Technology Management*. 23 (3), 171–175. doi:[10.1080/1097198X.2020.1794131](https://doi.org/10.1080/1097198X.2020.1794131).
- Lehmann, E.L. & Casella, G. (2006) *Theory of point estimation*. Springer Science & Business Media.
- Lehtonen, J. (1988) The information society and the new competence. *American Behavioral Scientist*. 32 (2), 104–111.
- Lenzerini, M. (2002) Data integration: A theoretical perspective. In: *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2002 pp. 233–246.
- Leski, J.M. (2014) Fuzzy $(c + p)$ k -means clustering and its application to a fuzzy rule-based classifier: Toward good generalization and good interpretability. *IEEE Transactions on Fuzzy Systems*. 23 (4), 802–812.
- Leskovec, J., Rajaraman, A. & Ullman, J.D. (2014) *Mining of Massive Datasets*, Cambridge University Press, Cambridge.
- Levendusky, M. (2009) *The partisan sort: How liberals became Democrats and conservatives became Republicans*. University of Chicago Press.
- Leviathan, Y. & Matias, Y. (2018) Google Duplex: An AI system for accomplishing real-world tasks over the phone. *Google AI blog*. 8. <http://research.google/pubs/google-duplex-an-ai-system-for-accomplishing-real-world-tasks-over-the-phone/>.
- Levinson, M.L. (2013) Education as a civic right: Using schools to challenge the civic empowerment gap. *Poverty & Race*.
- Levinson, M.L. (2010) *The civic empowerment gap: Defining the problem and locating solutions*.
- Levitsky, S. & Way, L.A. (2010) *Competitive authoritarianism: Hybrid regimes after the Cold War*. Cambridge University Press.

- Li, H., Yu, L. & He, W. (2019) The Impact of GDPR on Global Technology Development. *Journal of Global Information Technology Management*. 22 (1), 1–6. doi:[10.1080/1097198X.2019.1569186](https://doi.org/10.1080/1097198X.2019.1569186).
- Li, L., Chu, W., Langford, J. & Schapire, R.E. (2010) A contextual-bandit approach to personalized news article recommendation. In: *Proceedings of the 19th international conference on World wide web*. 2010 pp. 661–670.
- Li, Y. & Yang, T. (2018) Word Embedding for Understanding Natural Language: A Survey. In: S. Srinivasan (ed.). *Guide to Big Data Applications*. Studies in Big Data. Cham, Springer International Publishing. pp. 83–104. doi:[10.1007/978-3-319-53817-4_4](https://doi.org/10.1007/978-3-319-53817-4_4).
- Liebowitz, S.J. & Margolis, S.E. (1995) Path dependence, lock-in, and history. *The Journal of Law, Economics, and Organization*. 11 (1), 205–226.
- Lilkov, D. (2021) Regulating artificial intelligence in the EU: A risky game. *European View*. 20 (2), 166–174. doi:[10.1177/17816858211059248](https://doi.org/10.1177/17816858211059248).
- Ling, L.I. & Sheng-Hua, J.I.N. (2016) History of development and the newest reconstruction of Schwartz's theory of basic human values. *Journal of Psychological Science*. 39 (1), 191–199.
- Littman, M.L., Ajunwa, I., Berger, G., Boutilier, C., Currie, M., Doshi-Velez, F., Hadfield, G., Horowitz, M.C., Isbell, C. & Kitano, H. (2022) Gathering strength, gathering storms: The one hundred year study on artificial intelligence (AI100) 2021 study panel report. *arXiv preprint arXiv:2210.15767*.
- Liu, A., Wu, S. & Resnick, P. (2024) *How to Train Your YouTube Recommender to Avoid Unwanted Videos*. doi:[10.48550/arXiv.2307.14551](https://doi.org/10.48550/arXiv.2307.14551).
- Liu, T.-Y. (2009) Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*. 3 (3), 225–331.
- Liu, Y.C. & Huang, M.Q. (2021) Examining the Matthew Effect on YouTube Recommendation System. In: *2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. 18 November 2021 pp. 146–148. doi:[10.1109/TAAI54685.2021.00035](https://doi.org/10.1109/TAAI54685.2021.00035).
- Lohr, S.L. (2021) *Sampling: design and analysis*. Chapman and Hall/CRC.

<https://www.taylorfrancis.com/books/mono/10.1201/9780429298899/sampling-sharon-lohr>.

- Lohtia, R., Donthu, N. & Hershberger, E.K. (2003) The impact of content and design elements on banner advertising click-through rates. *Journal of advertising Research*. 43 (4), 410–418.
- Lokot, T. (2020) Data subjects vs. people's data: Competing discourses of privacy and power in modern Russia. *Media and Communication*. 8 (2), 314–322.
- Lops, P., De Gemmis, M. & Semeraro, G. (2011) Content-based recommender systems: State of the art and trends. *Recommender systems handbook*. 73–105.
- Lu, K., Cai, X., Ajiferuke, I. & Wolfram, D. (2017) Vocabulary size and its effect on topic representation. *Information Processing & Management*. 53 (3), 653–665.
- Lu, Y. (2019) Artificial intelligence: a survey on evolution, models, applications and future trends. *Journal of Management Analytics*. 6 (1), 1–29. doi:[10.1080/23270012.2019.1570365](https://doi.org/10.1080/23270012.2019.1570365).
- Lumbut, C. & Ponnoprat, D. (2025) Investigating privacy leakage in dimensionality reduction methods via reconstruction attack. *Journal of Information Security and Applications*. 92, 104102.
- Lysaght, T. (2022) Anticipatory Governance and Foresight in Regulating for Uncertainty. *The American Journal of Bioethics*. 22 (1), 51–53. doi:[10.1080/15265161.2021.2001111](https://doi.org/10.1080/15265161.2021.2001111).
- MacCarthy, M. (2017) Standards of fairness for disparate impact assessment of big data algorithms. *Cumb. L. Rev.* 48, 67.
- Mahakal, P. (2025) *Model VS AI Agents*. 17 April 2025. Medium. <https://levelup.gitconnected.com/model-vs-ai-agents-9e5966d346a3> [Accessed: 19 April 2025].
- Mahadevan, A. & Mathioudakis, M. (2024) Cost-aware retraining for machine learning. *Knowledge-Based Systems*. 293, 111610.
- Mahoney, J. (2000) Path dependence in historical sociology. *Theory and society*. 29 (4), 507–548.
- Makhortykh, M. & Bastian, M. (2022) Personalizing the war: Perspectives for the adoption of news recommendation algorithms in the media coverage of the conflict in Eastern Ukraine. *Media, War & Conflict*. 15 (1), 25–45. doi:[10.1177/1750635220906254](https://doi.org/10.1177/1750635220906254).

- Makridakis, S. (2017) The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*. 90, 46–60. doi:[10.1016/j.futures.2017.03.006](https://doi.org/10.1016/j.futures.2017.03.006).
- Malgieri, G. & Custers, B. (2018) Pricing privacy—the right to know the value of your personal data. *Computer Law & Security Review*. 34 (2), 289–303.
- Manrai, L.A. & Manrai, A.K. (2011) Hofstede s cultural Dimensions and tourist behaviors: A review and conceptual Framework. *Journal of Economics, Finance and Administrative Science*. 16 (31), 23–48.
- Mansell, R. & Raboy, M. (2011) *The Handbook of Global Media and Communication Policy*. 1st edition. Wiley. doi:[10.1002/9781444395433](https://doi.org/10.1002/9781444395433).
- Mansoorizadeh, M., Shahidi, M.-M. & Nazari, A. (2023) *A Framework for Accurate Recommendations and Explanation Generation Using Multi-Task Learning*. <https://www.researchsquare.com/article/rs-3710935/latest>.
- Mantelero, A. (2018) AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*. 34 (4), 754–772.
- Maréchal, N. (2017) Networked authoritarianism and the geopolitics of information: Understanding Russian Internet policy. *Media and communication*. 5 (1), 29–41.
- Marchant, G.E. (2011) The growing gap between emerging technologies and the law. *The growing gap between emerging technologies and legal-ethical oversight: The pacing problem*. 19–33.
- Marianna, M. (2014) Traditional values and modern families: legal understanding of tradition andModernity in contemporary Russia. *Журнал исследований социальной политики*. 12 (4), 625–638.
- Markl, N. (2023) *Language variation, automatic speech recognition and algorithmic bias*. <https://era.ed.ac.uk/handle/1842/41277>.
- Marr, B. (2019) *The 10 Best Examples Of How AI Is Already Used In Our Everyday Life*. 2019. Forbes. <https://www.forbes.com/sites/bernardmarr/2019/12/16/the-10-best-examples-of-how-ai-is-already-used-in-our-everyday-life/> [Accessed: 7 August 2025].

- Massanari, A. (2017) #Gamergate and The Fapping: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*. 19 (3), 329–346. doi:[10.1177/1461444815608807](https://doi.org/10.1177/1461444815608807).
- Massey, R. & Machin, E. (2023) *Private capital: Impact of AI and its regulation for investors in 2024 | Insights | Ropes & Gray LLP*. 2023. <https://www.ropesgray.com/en/insights/viewpoints/102ivhz/private-capital-impact-of-ai-and-its-regulation-for-investors-in-2024> [Accessed: 9 July 2025].
- Masters, B. (2025) Tennessee drops BlackRock ESG suit after asset manager agrees concessions. *Financial Times*. 17 January. <https://www.ft.com/content/6dbd5600-820a-4899-8a8b-0b276e634a0a>.
- Mayer, J.R. & Mitchell, J.C. (2012) Third-party web tracking: Policy and technology. In: *2012 IEEE symposium on security and privacy*. 2012 IEEE. pp. 413–427. https://ieeexplore.ieee.org/abstract/document/6234427/?casa_token=YJHQTRb9ERMAAAAA:DBdOb9GwUNkH1FzrIW049hMvlobptQTY0Ue7xjBuyMJxXyaYl5jEiuOMlcmCjK2MChmEjQzl8B0.
- McGinnis, J.O. (2012) *Accelerating Democracy*. In: *Accelerating Democracy*. Princeton University Press. p.
- McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., et al. (2020) International evaluation of an AI system for breast cancer screening. *Nature*. 577 (7788), 89–94. doi:[10.1038/s41586-019-1799-6](https://doi.org/10.1038/s41586-019-1799-6).
- McNee, S.M., Riedl, J. & Konstan, J.A. (2006) Being accurate is not enough: how accuracy metrics have hurt recommender systems. In: *CHI'06 extended abstracts on Human factors in computing systems*. 2006 pp. 1097–1101.
- Medhat, W., Hassan, A. & Korashy, H. (2014) Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*. 5 (4), 1093–1113.
- Meduza (2024a) «Война — не игра для женщин»: российская команда создала в Германии документальный спектакль о беженках Он критикует не только войну и гендерное неравенство, но и сам театр. 2024. Meduza. <https://meduza.io/feature/2024/04/06/voyna-ne-igra-dlya-zhenschin-rossiyskaya-komanda-sozdala-v-germanii-dokumentalnyy-spektakl-o-bezhenkah> [Accessed: 5 January 2025].

- Meduza (2024b) *Мишустин заявил, что в искусственный интеллект российской и зарубежной разработки «заложены разные картины мира»*. 2024. Meduza. <https://meduza.io/news/2024/02/02/mishustin-zayavil-chto-v-iskusstvennyy-intellekt-rossiyskoy-i-zarubezhnoy-razrabotki-zalozheny-raznye-kartiny-mira> [Accessed: 15 July 2025].
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2021) A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*. 54 (6), 1–35.
- Melkozorova, V. (2024) *Russia to prepare for war with NATO*. 16 December 2024. POLITICO. <https://www.politico.eu/article/russia-to-prepare-for-war-with-nato-andrei-belousov-vladimir-putin/> [Accessed: 5 January 2025].
- Melville, P. & Sindhvani, V. (2010) Recommender systems. *Encyclopedia of machine learning*. 1, 829–838.
- Mesbahi, M. & Egerstedt, M. (2010) *Graph theoretic methods in multiagent networks*. <https://www.torrossa.com/gs/resourceProxy?an=5573598&publisher=FZO137>.
- Metropolis, N. & Ulam, S. (1949) The Monte Carlo Method. *Journal of the American Statistical Association*. 44 (247), 335–341. doi:[10.1080/01621459.1949.10483310](https://doi.org/10.1080/01621459.1949.10483310).
- Milano, S., Taddeo, M. & Floridi, L. (2020) Recommender systems and their ethical challenges. *Ai & Society*. 35, 957–967.
- Mili, A. & Sheldon, F. (2009) Challenging the mean time to failure: Measuring dependability as a mean failure cost. In: *2009 42nd Hawaii International Conference on System Sciences*. 2009 IEEE. pp. 1–10. <https://ieeexplore.ieee.org/abstract/document/4755411/>.
- Miller, E., Manz, E., Russ, R., Stroupe, D. & Berland, L. (2018) Addressing the epistemic elephant in the room: Epistemic agency and the next generation science standards. *Journal of Research in Science Teaching*. 55 (7), 1053–1075. doi:[10.1002/tea.21459](https://doi.org/10.1002/tea.21459).
- Mills, M. & Keddie, A. (2010) Cultural reductionism and the media: polarising discourses around schools, violence and masculinity in an age of terror. *Oxford Review of Education*. 36 (4), 427–444. doi:[10.1080/03054985.2010.494449](https://doi.org/10.1080/03054985.2010.494449).

- Milmo, D. (2021) Facebook to shut facial recognition system and delete 1bn 'faceprints'. *The Guardian*. 2 November.
<https://www.theguardian.com/technology/2021/nov/02/facebook-to-shut-facial-recognition-system-and-delete-1bn-faceprints>.
- Milmo, D. (2023) Sarah Silverman sues OpenAI and Meta claiming AI training infringed copyright. *The Guardian*. 10 July.
<https://www.theguardian.com/technology/2023/jul/10/sarah-silverman-sues-openai-meta-copyright-infringement>.
- Mitchell, M., Baker, D., Moorosi, N., Denton, E., Hutchinson, B., Hanna, A., Gebru, T. & Morgenstern, J. (2020) Diversity and Inclusion Metrics in Subset Selection. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 7 February 2020 New York NY USA, ACM. pp. 117–123. doi:[10.1145/3375627.3375832](https://doi.org/10.1145/3375627.3375832).
- Mittelstadt, B. (2019) Principles alone cannot guarantee ethical AI. *Nature machine intelligence*. 1 (11), 501–507.
- Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society*. 3 (2), 2053951716679679.
- Mizgajski, J. & Morzy, M. (2019) Affective recommender systems in online news industry: how emotions influence reading choices. *User Modeling and User-Adapted Interaction*. 29 (2), 345–379.
doi:[10.1007/s11257-018-9213-x](https://doi.org/10.1007/s11257-018-9213-x).
- Mo, C., Jiang, C., Lei, X., Lai, S., Deng, Y., Cen, W., Sun, G. & Xing, Z. (2022) Combining Standard Artificial Intelligence Models, Pre-Processing Techniques, and Post-Processing Methods to Improve the Accuracy of Monthly Runoff Predictions in Karst-Area Watersheds. *Applied Sciences*. 13 (1), 88.
- Modesto De Araujo, L.P., Teles De Oliveira, C.L., Da Hora Rodrigues, K.R. & Garcia Manzato, M. (2022) Empirical Studies Aimed at Understanding Conversational Recommender Systems and Accessibility Aspects. In: V.G. Duffy, Q. Gao, J. Zhou, M. Antona, & C. Stephanidis (eds). *HCI International 2022 – Late Breaking Papers: HCI for Health, Well-being, Universal Access and Healthy Aging*. Lecture Notes in Computer Science. Cham, Springer Nature Switzerland. pp. 462–478. doi:[10.1007/978-3-031-17902-0_33](https://doi.org/10.1007/978-3-031-17902-0_33).

- Molitor, D., Needell, D. & Ward, R. (2021) Bias of homotopic gradient descent for the hinge loss. *Applied Mathematics & Optimization*. 84, 621–647.
- Möller, J., Trilling, D., Helberger, N. & van Es, B. (2020) Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. In: *Digital media, political polarization and challenges to democracy*. Routledge. pp. 45–63.
<https://www.taylorfrancis.com/chapters/edit/10.4324/9780429243912-4/blame-algorithm-judith-m%C3%B6ller-damian-trilling-natali-helberger-bram-van-es>.
- Montanaro, B., Croce, A. & Ughetto, E. (2024) Venture capital investments in artificial intelligence. *Journal of Evolutionary Economics*. 34 (1), 1–28. doi:[10.1007/s00191-024-00857-7](https://doi.org/10.1007/s00191-024-00857-7).
- Montgomery, B. (2024) Apple delays launch of AI-powered features in Europe, blaming EU rules. *The Guardian*. 21 June.
<https://www.theguardian.com/technology/article/2024/jun/21/apple-ai-europe-regulation>.
- Morgan, D. (2023) Anticipatory regulatory instruments for AI systems: A comparative study of regulatory sandbox schemes. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 8 August 2023 Montréal QC Canada, ACM. pp. 980–981.
doi:[10.1145/3600211.3604732](https://doi.org/10.1145/3600211.3604732).
- Morgus, R. (2019) The spread of Russia’s digital Authoritarianism. *Artificial Intelligence, China, Russia, and the Global Order*. 89–97.
- Morozov, E. (2016) 76. The Net Delusion: How Not to Liberate the World. In: R. Blaug & J. Schwarzmantel (eds). *Democracy*. Columbia University Press. pp. 436–440. doi:[10.7312/blau17412-092](https://doi.org/10.7312/blau17412-092).
- Morozov, E. (2013) *To save everything, click here: The folly of technological solutionism*. PublicAffairs.
- Mueller, G.H. (1979) The notion of rationality in the work of Max Weber. *European Journal of Sociology/Archives Européennes de Sociologie*. 20 (1), 149–171.
- Mulgan, G.J. (1991) *Communication and control: Networks and the new economies of communication*. Guilford Publications, Inc.

- Munger, K. & Phillips, J. (2022) Right-Wing YouTube: A Supply and Demand Perspective. *The International Journal of Press/Politics*. 27 (1), 186–219. doi:[10.1177/1940161220964767](https://doi.org/10.1177/1940161220964767).
- Murphy, J.M., Berwick, D.M., Weinstein, M.C., Borus, J.F., Budman, S.H. & Klerman, G.L. (1987) Performance of screening and diagnostic tests: Application of receiver operating characteristic analysis. *Archives of general psychiatry*. 44 (6), 550–555.
- Murthy, M.N. (1967) *Sampling theory and methods*.
<https://www.cabidigitallibrary.org/doi/full/10.5555/19702700466>.
- Musick, G., Duan, W., Najafian, S., Sengupta, S., Flathmann, C., Knijnenburg, B. & McNeese, N. (2024) To Share or Not to Share: Understanding and Modeling Individual Disclosure Preferences in Recommender Systems for the Workplace. *Proceedings of the ACM on Human-Computer Interaction*. 8 (GROUP), 1–28. doi:[10.1145/3633074](https://doi.org/10.1145/3633074).
- Musosoup (2025) *The Sounds of Spotify: Exploring Genres, Playlists, and Discovery Features*. 2025. Musosoup.
<https://musosoup.com/blog/the-sounds-of-spotify> [Accessed: 10 August 2025].
- Muthukumar, V., Narang, A., Subramanian, V., Belkin, M., Hsu, D. & Sahai, A. (2021) Classification vs regression in overparameterized regimes: Does the loss function matter? *The Journal of Machine Learning Research*. 22 (1), 10104–10172.
- Mutz, D.C. (2008) Is deliberative democracy a falsifiable theory? *Annual Review of Political Science*. 11, 521.
- Myhre, M.H., Aasland, A. & Holm-Hansen, J. (2024) ‘Crimea will forever be Russian’: dissenting Norwegian media discourses on Russia’s annexation of Crimea. *European Politics and Society*. 25 (1), 185–208. doi:[10.1080/23745118.2022.2103286](https://doi.org/10.1080/23745118.2022.2103286).
- Namey, E., Guest, G., Thairu, L. & Johnson, L. (2008) Data reduction techniques for large qualitative data sets. *Handbook for team-based qualitative research*. 2 (1), 137–161.
- Narayanan, A. & Shmatikov, V. (2008) Robust de-anonymization of large sparse datasets. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 2008 IEEE. pp. 111–125.
<https://ieeexplore.ieee.org/abstract/document/4531148/>.

- Nasr, M., Songi, S., Thakurta, A., Papernot, N. & Carlin, N. (2021) Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning. In: *2021 IEEE Symposium on Security and Privacy (SP)*. May 2021 pp. 866–882. doi:[10.1109/SP40001.2021.00069](https://doi.org/10.1109/SP40001.2021.00069).
- Nation, I. (2006) How Large a Vocabulary is Needed For Reading and Listening? *The Canadian Modern Language Review*. 63 (1), 59–82. doi:[10.3138/cmlr.63.1.59](https://doi.org/10.3138/cmlr.63.1.59).
- Nedobity, W. (1985) Terminology and artificial intelligence. *KNOWLEDGE ORGANIZATION*. 12 (1), 17–19. doi:[10.5771/0943-7444-1985-1-17](https://doi.org/10.5771/0943-7444-1985-1-17).
- Neisser, U. & Jopling, D.A. (1997) *The conceptual self in context: culture experience self understanding*. Cambridge University Press.
- Nemitz, P. (2018) Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 376 (2133), 20180089. doi:[10.1098/rsta.2018.0089](https://doi.org/10.1098/rsta.2018.0089).
- von Neumann, J. & Ulam, S. (2024) Monte Carlo Simulation. *Gateway to Business Analytics*. 37.
- NEUROCLIMA (2025) News - NEUROCLIMA. 2025. <https://neuroclima.eu/category/news/>.
- Nguyen, C.T. (2020) Echo chambers and epistemic bubbles. *Episteme*. 17 (2), 141–161.
- Nguyen, T.T., Hui, P.-M., Harper, F.M., Terveen, L. & Konstan, J.A. (2014) Exploring the filter bubble: the effect of using recommender systems on content diversity. In: *Proceedings of the 23rd international conference on World wide web*. WWW '14. 7 April 2014 New York, NY, USA, Association for Computing Machinery. pp. 677–686. doi:[10.1145/2566486.2568012](https://doi.org/10.1145/2566486.2568012).
- Nguyen, T.T., Maxwell Harper, F., Terveen, L. & Konstan, J.A. (2018) User personality and user satisfaction with recommender systems. *Information systems frontiers*. 20, 1173–1189.
- Nguyen-Thanh, N., Marinca, D., Khawam, K., Rohde, D., Vasile, F., Lohan, E.S., Martin, S. & Quadri, D. (2019) Recommendation system-based upper confidence bound for online advertising. *arXiv preprint arXiv:1909.04190*.

- Nichols, J. (2018) *Linguistic diversity in space and time*. University of Chicago Press.
- Nie, N.H., Junn, J. & Stehlik-Barry, K. (1996) *Education and democratic citizenship in America*. University of Chicago Press.
- Nisar, T.M. & Prabhakar, G. (2017) What factors determine e-satisfaction and consumer spending in e-commerce retailing? *Journal of retailing and consumer services*. 39, 135–144.
- Nitsevich, V.F., Moiseev, V.V., Stroev, V.V. & Sudorgin, O.A. (2020) Actual Problems of Business in Russia. In: D.B. Solovev (ed.). *Smart Technologies and Innovations in Design for Control of Technological Processes and Objects: Economy and Production*. Smart Innovation, Systems and Technologies. Cham, Springer International Publishing. pp. 815–825. doi:[10.1007/978-3-030-15577-3_75](https://doi.org/10.1007/978-3-030-15577-3_75).
- Noble, D.F. (1978) Social Choice in Machine Design: The Case of Automatically Controlled Machine Tools, and a Challenge for Labor. *Politics & Society*. 8 (3–4), 313–347. doi:[10.1177/003232927800800302](https://doi.org/10.1177/003232927800800302).
- Novelli, C., Taddeo, M. & Floridi, L. (2024) Accountability in artificial intelligence: what it is and how it works. *AI & SOCIETY*. 39 (4), 1871–1882. doi:[10.1007/s00146-023-01635-y](https://doi.org/10.1007/s00146-023-01635-y).
- Ntoutsis, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdli, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S. & Krasanakis, E. (2020) Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 10 (3), e1356.
- Numan, G. (2020) Testing artificial intelligence. In: *The Future of Software Quality Assurance*. Springer, Cham. pp. 123–136.
- O'Connor, C. & Joffe, H. (2020) Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods*. 19, 160940691989922. doi:[10.1177/1609406919899220](https://doi.org/10.1177/1609406919899220).
- OECD (2013) *Government at a Glance 2013*. 14 November 2013. OECD. https://www.oecd.org/en/publications/government-at-a-glance-2013_gov_glance-2013-en.html [Accessed: 19 August 2025].

- Oleinik, A. (2024) Putting Discourses About the War in Ukraine on a Map: How Different is Everyone's Story? *International Politics*. doi:10.1057/s41311-024-00564-x.
- Olson, H.A. (2013) *The power to name: locating the limits of subject representation in libraries*. Springer Science & Business Media.
- O'Mahony, M., Hurley, N., Kushmerick, N. & Silvestre, G. (2004) Collaborative recommendation: A robustness analysis. *ACM Transactions on Internet Technology (TOIT)*. 4 (4), 344–377.
- ORBIS (2025) *Homepage*. 9 June 2025. Orbis Project. <https://orbis-project.eu/> [Accessed: 19 August 2025].
- Orlova, A.V. (2019) 'Digital Sovereignty,' Anonymity and Freedom of Expression: Russia's Fight to Re-Shape Internet Governance. *UC Davis J. Int'l L. & Pol'y*. 26, 225.
- Orobchuk, D. (2024) Charting language shift through Ukraine's social media actors. *Canadian Slavonic Papers*. 66 (3–4), 431–455. doi:10.1080/00085006.2024.2415177.
- Ortega, F., Sanchez, J.-L., Bobadilla, J. & Gutierrez, A. (2013) Improving collaborative filtering-based recommender systems results using Pareto dominance. *Information Sciences*. 239, 50–61.
- Osborne, J. (2010) Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation*. 15 (1), 12.
- Osipov, A. (2010) National cultural autonomy in Russia: A case of symbolic law. *Review of Central and East European Law*. 35 (1), 27–57.
- Ovadya, A. (2022) *Bridging-Based Ranking*.
- Page, S.E. (2006) Path dependence. *Quarterly Journal of Political Science*. 1 (1), 87–115.
- Päiväranta, T. & Sæbø, Ø. (2006) Models of e-democracy. *Communications of the Association for Information Systems*. 17 (1), 37.
- Pan, S.J. & Yang, Q. (2010) A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*. 22 (10), 1345–1359.
- Pan, Y., He, F. & Yu, H. (2020) Learning social representations with deep autoencoder for recommender system. *World Wide Web*. 23, 2259–2279.
- Papadogiannakis, E., Papadopoulos, P., Kourtellis, N. & Markatos, E.P. (2021) User Tracking in the Post-cookie Era: How Websites Bypass

GDPR Consent to Track Users. In: *Proceedings of the Web Conference 2021*. 19 April 2021 Ljubljana Slovenia, ACM. pp. 2130–2141.
doi:[10.1145/3442381.3450056](https://doi.org/10.1145/3442381.3450056).

- Papadopoulos, P., Kourtellis, N. & Markatos, E. (2019) Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask. In: *The World Wide Web Conference*. 13 May 2019 San Francisco CA USA, ACM. pp. 1432–1442.
doi:[10.1145/3308558.3313542](https://doi.org/10.1145/3308558.3313542).
- Papyshev, G. & Yarime, M. (2023) The challenges of industry self-regulation of AI in emerging economies: implications of the case of Russia for public policy and institutional development. In: *Elgar companion to regulating AI and big data in emerging economies*. Edward Elgar Publishing. pp. 81–98.
<https://www.elgaronline.com/abstract/book/9781785362408/chapter4.xml>.
- Paraschakis, D. (2017) Towards an ethical recommendation framework. In: *2017 11th international conference on research challenges in information science (RCIS)*. 2017 IEEE. pp. 211–220.
https://ieeexplore.ieee.org/abstract/document/7956539/?casa_token=jAGK-RxceG8AAAAA:o159VP-0qrIGxclN0QB9Llp0wCDBXZj4G64BsriXB5hXWcA8hZs1YeDB8GzLURV2AhBlmF1MA.
- Pariser, E. (2011) *The filter bubble: What the Internet is hiding from you*. penguin UK.
- Parker, K. (2019) The growing partisan divide in views of higher education. *Pew Research Center*. 19, 2–3.
- Pasitselska, O. (2017) Ukrainian crisis through the lens of Russian media: Construction of ideological discourse. *Discourse & Communication*. 11 (6), 591–609. doi:[10.1177/1750481317714127](https://doi.org/10.1177/1750481317714127).
- Patel, C., Patel, A. & Patel, D. (2012) Optical character recognition by open source OCR tool tesseract: A case study. *International journal of computer applications*. 55 (10).
<https://www.academia.edu/download/100190454/3e47cc647c47a1a249e1103047dd5b002b5a.pdf>.
- Paterson, G. (2024) Sock Puppetry in Online Communication. *Philosophy*. 99 (3), 461–478.

- Pathe, S., Krieg, G. & Bradner, E. (2024) *Takeaways from Super Tuesday* | *CNN Politics*. 6 March 2024. CNN.
<https://www.cnn.com/2024/03/06/politics/takeaways-super-tuesday/index.html> [Accessed: 1 January 2025].
- Paxton, P. (2002) Social capital and democracy: An interdependent relationship. *American sociological review*. 254–277.
- Pazzani, M.J. & Billsus, D. (2007) Content-based recommendation systems. *The adaptive web: methods and strategies of web personalization*. 325–341.
- Pedersen, C.L. (2019) Is Silicon Valley Diverse Enough to Resist Reduction? *Journal of Design and Science*.
<https://assets.pubpub.org/4fnns6ut/5aacb543-3dbd-45b3-91e6-34bef69570ba.pdf>.
- Pendlebury, F., Pierazzi, F., Jordaney, R., Kinder, J. & Cavallaro, L. (2019) {TESSERACT}: Eliminating experimental bias in malware classification across space and time. In: *28th USENIX security symposium (USENIX Security 19)*. 2019 pp. 729–746.
<https://www.usenix.org/conference/usenixsecurity19/presentation/pendlebury>.
- Perez, O. & Wimer, N. (2023) Algorithmic Constitutionalism. *Ind. J. Global Legal Stud.* 30, 81.
- Perrigo, B. (2023) *Exclusive: The \$2 Per Hour Workers Who Made ChatGPT Safer*. 18 January 2023. Time.
<https://time.com/6247678/openai-chatgpt-kenya-workers/> [Accessed: 21 July 2025].
- Phothongsunan, S. (2010) *Interpretive paradigm in educational research*.
<https://repository.au.edu/bitstream/6623004553/13708/1/galaxy-iele-v2-n1-1-oct-10.pdf>.
- Pi, Q., Bian, W., Zhou, G., Zhu, X. & Gai, K. (2019) Practice on Long Sequential User Behavior Modeling for Click-Through Rate Prediction. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 25 July 2019 pp. 2671–2679.
doi:[10.1145/3292500.3330666](https://doi.org/10.1145/3292500.3330666).
- Pierson, P. (2000) Increasing returns, path dependence, and the study of politics. *American political science review*. 94 (2), 251–267.

- Pigman, L. (2019) Russia's vision of cyberspace: a danger to regime security, public safety, and societal norms and cohesion. *Journal of Cyber Policy*. 4 (1), 22–34. doi:[10.1080/23738871.2018.1546884](https://doi.org/10.1080/23738871.2018.1546884).
- Pintor, M., Demetrio, L., Sotgiu, A., Demontis, A., Carlini, N., Biggio, B. & Roli, F. (2022) Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples. *Advances in Neural Information Processing Systems*. 35, 23063–23076.
- Plano Clark, V.L. (2017) Mixed methods research. *The Journal of Positive Psychology*. 12 (3), 305–306. doi:[10.1080/17439760.2016.1262619](https://doi.org/10.1080/17439760.2016.1262619).
- Po'ttker, H. (2003) News and its communicative quality: the inverted pyramid—when and why did it appear? *Journalism Studies*. 4 (4), 501–511. doi:[10.1080/1461670032000136596](https://doi.org/10.1080/1461670032000136596).
- Pohlhaus, G. (2020) Epistemic Agency Under Oppression. *Philosophical Papers*. 49 (2), 233–251. doi:[10.1080/05568641.2020.1780149](https://doi.org/10.1080/05568641.2020.1780149).
- Polyakova, A. & Meserole, C. (2019) Exporting digital authoritarianism: The Russian and Chinese models. *Policy brief, democracy and disorder series*. 1–22.
- Ponti, M., Kittler, J., Riva, M., de Campos, T. & Zor, C. (2017) A decision cognizant Kullback–Leibler divergence. *Pattern Recognition*. 61, 470–478.
- Porcaro, L., Vinagre, J., Frau, P., Hupont, I. & Gómez, E. (2023) *Behind Recommender Systems: the Geography of the ACM RecSys Community*. doi:[10.48550/arXiv.2309.03512](https://doi.org/10.48550/arXiv.2309.03512).
- Poster, M. (1990) *The mode of information: Poststructuralism and social context*. University of Chicago Press.
- Pöttker, H. (2005) The News Pyramid and its Origin from the American Journalism in the 19th Century. *Høyer/Pöttker (2005)*. 51–64.
- Preil, D. & Krapp, M. (2022) Artificial intelligence-based inventory management: a Monte Carlo tree search approach. *Annals of Operations Research*. 308 (1–2), 415–439. doi:[10.1007/s10479-021-03935-2](https://doi.org/10.1007/s10479-021-03935-2).
- Price, W. & Nicholson, I.I. (2019) Medical AI and contextual bias. *Harv. JL & Tech*. 33, 65.

- Priem, J., Piwowar, H. & Orr, R. (2022) *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. doi:[10.48550/arXiv.2205.01833](https://doi.org/10.48550/arXiv.2205.01833).
- Priest, G.L. (1986) The current insurance crisis and modern tort law. *Yale Lj.* 96, 1521.
- Prokopenko, A. (2024) *Russia's Economic Gamble: The Hidden Costs of War-Driven Growth* | Carnegie Endowment for International Peace. 2024. [https://carnegieendowment.org/russia-
eurasia/politika/2024/12/russia-economy-difficulties?lang=en](https://carnegieendowment.org/russia-eurasia/politika/2024/12/russia-economy-difficulties?lang=en) [Accessed: 5 January 2025].
- Proudfoot, D. (2011) Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence.* 175 (5–6), 950–957.
- Prunkl, C., Ashurst, C., Anderljung, M., Webb, H., Leike, J. & Dafoe, A. (2021) Institutionalising Ethics in AI through Broader Impact Requirements. *Nature Machine Intelligence.* 3 (2), 104–110. doi:[10.1038/s42256-021-00298-y](https://doi.org/10.1038/s42256-021-00298-y).
- Psaros, A.F., Kawaguchi, K. & Karniadakis, G.E. (2022) Meta-learning PINN loss functions. *Journal of computational physics.* 458, 111121.
- Pu, P., Chen, L. & Hu, R. (2011) A user-centric evaluation framework for recommender systems. In: *Proceedings of the fifth ACM conference on Recommender systems*. 2011 pp. 157–164.
- Pu, P., Chen, L. & Hu, R. (2012) Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction.* 22 (4–5), 317–355. doi:[10.1007/s11257-011-
9115-7](https://doi.org/10.1007/s11257-011-9115-7).
- Puhakainen, E. & Väyrynen, K.E. (2021) *The benefits and challenges of technology neutral regulation—a scoping review*.
- Pujol, E., Hohlfeld, O. & Feldmann, A. (2015) Annoyed Users: Ads and Ad-Block Usage in the Wild. In: *Proceedings of the 2015 Internet Measurement Conference*. 28 October 2015 Tokyo Japan, ACM. pp. 93–106. doi:[10.1145/2815675.2815705](https://doi.org/10.1145/2815675.2815705).
- Qadri, R., Diaz, M., Wang, D. & Madaio, M. (2025) *The Case for 'Thick Evaluations' of Cultural Representation in AI*. doi:[10.48550/arXiv.2503.19075](https://doi.org/10.48550/arXiv.2503.19075).

- Qin, J., Zhang, W., Wu, X., Jin, J., Fang, Y. & Yu, Y. (2020) User Behavior Retrieval for Click-Through Rate Prediction. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 25 July 2020 pp. 2347–2356. doi:[10.1145/3397271.3401440](https://doi.org/10.1145/3397271.3401440).
- Qiu, G., He, X., Zhang, F., Shi, Y., Bu, J. & Chen, C. (2010) DASA: Dissatisfaction-oriented Advertising based on Sentiment Analysis. *Expert Systems with Applications*. 37 (9), 6182–6191. doi:[10.1016/j.eswa.2010.02.109](https://doi.org/10.1016/j.eswa.2010.02.109).
- Radaelli, C.M. (2005) *What does regulatory impact assessment mean in Europe?* https://ore.exeter.ac.uk/repository/bitstream/handle/10036/22912/Radaelli%20regulatory_impact_assessment_Europe.pdf?sequence=1.
- Ragin, C.C. (2014) *The comparative method: Moving beyond qualitative and quantitative strategies*. Univ of California Press.
- Raimondi, F., Skene, J. & Emmerich, W. (2008) Efficient online monitoring of web-service SLAs. In: *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering*. 9 November 2008 Atlanta Georgia, ACM. pp. 170–180. doi:[10.1145/1453101.1453125](https://doi.org/10.1145/1453101.1453125).
- Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. & Barnes, P. (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 27 January 2020 Barcelona Spain, ACM. pp. 33–44. doi:[10.1145/3351095.3372873](https://doi.org/10.1145/3351095.3372873).
- Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., Benítez, J.M. & Herrera, F. (2016) Data discretization: taxonomy and big data challenge. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 6 (1), 5–21.
- Raza, S. & Ding, C. (2022) News recommender system: a review of recent progress, challenges, and opportunities. *Artificial Intelligence Review*. 55 (1), 749–800. doi:[10.1007/s10462-021-10043-x](https://doi.org/10.1007/s10462-021-10043-x).
- Rep. Johnson, E.B. [D-T.-30 (2020) *H.R.6216 - 116th Congress (2019-2020): National Artificial Intelligence Initiative Act of 2020*. 3 December

2020. <https://www.congress.gov/bill/116th-congress/house-bill/6216> [Accessed: 19 July 2025].

- Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. & Riedl, J. (1994) Grouplens: An open architecture for collaborative filtering of netnews. In: *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. 1994 pp. 175–186.
- Resnick, P. & Varian, H.R. (1997) Recommender systems. *Communications of the ACM*. 40 (3), 56–58.
- Ribeiro, M.H., Ottoni, R., West, R., Almeida, V.A.F. & Meira, W. (2020) Auditing radicalization pathways on YouTube. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. 27 January 2020 New York, NY, USA, Association for Computing Machinery. pp. 131–141. doi:[10.1145/3351095.3372879](https://doi.org/10.1145/3351095.3372879).
- Ricci, F. (2014) *Recommender Systems: Models and Techniques*. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5d64e815df846fa3a10e3274cd33fe1a2d94f9f1>.
- F. Ricci, L. Rokach, & B. Shapira (eds) (2015) *Recommender Systems Handbook*. Boston, MA, Springer US. doi:[10.1007/978-1-4899-7637-6](https://doi.org/10.1007/978-1-4899-7637-6).
- F. Ricci, L. Rokach, & B. Shapira (eds) (2022) *Recommender Systems Handbook*. New York, NY, Springer US. doi:[10.1007/978-1-0716-2197-4](https://doi.org/10.1007/978-1-0716-2197-4).
- Richards, B. (1987) Type/token ratios: What do they really tell us? *Journal of child language*. 14 (2), 201–209.
- Ritzer, G. (2015) Prosumer capitalism. *The Sociological Quarterly*. 56 (3), 413–445.
- Ritzer, G., Dean, P. & Jurgenson, N. (2012) The coming of age of the prosumer. *American behavioral scientist*. 56 (4), 379–398.
- Robins, K. & Webster, F. (2003) *Times of the technoculture: From the information society to the virtual life*. Routledge.
- Robins-Early, N. (2024) Google restricts AI chatbot Gemini from answering questions on 2024 elections. *The Guardian*. 12 March. <https://www.theguardian.com/us-news/2024/mar/12/google-ai-gemini-2024-election>.
- Rodrigues, R. (2020) Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. *Journal of Responsible Technology*. 4, 100005.

- Roh, Y., Lee, K., Whang, S.E. & Suh, C. (2021) *FairBatch: Batch Selection for Model Fairness*. doi:[10.48550/arXiv.2012.01696](https://doi.org/10.48550/arXiv.2012.01696).
- Roose, K. (2023) A Conversation With Bing's Chatbot Left Me Deeply Unsettled. *The New York Times*. 16 February.
<https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html>.
- Roose, K. (2019) The Making of a YouTube Radical. *The New York Times*. 8 June.
<https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html>.
- Rosasco, L., De Vito, E., Caponnetto, A., Piana, M. & Verri, A. (2004) Are loss functions all the same? *Neural computation*. 16 (5), 1063–1076.
- Roselli, D., Matthews, J. & Talagala, N. (2019) Managing bias in AI. In: *Companion Proceedings of The 2019 World Wide Web Conference*. 2019 pp. 539–544.
- Roshan, K. & Zafar, A. (2021) Utilizing XAI technique to improve autoencoder based model for computer network anomaly detection with shapley additive explanation(SHAP). *International journal of Computer Networks & Communications*. 13 (6), 109–128.
doi:[10.5121/ijcnc.2021.13607](https://doi.org/10.5121/ijcnc.2021.13607).
- Ross, E.A. & Gross, M. (2017) *Social control: A survey of the foundations of order*. Routledge.
- Rotolo, D., Rafols, I., Hopkins, M.M. & Leydesdorff, L. (2017) Strategic intelligence on emerging technologies: Scientometric overlay mapping. *Journal of the Association for Information Science and Technology*. 68 (1), 214–233. doi:[10.1002/asi.23631](https://doi.org/10.1002/asi.23631).
- Roy, D. & Dutta, M. (2022) A systematic review and research perspective on recommender systems. *Journal of Big Data*. 9 (1), 59.
doi:[10.1186/s40537-022-00592-5](https://doi.org/10.1186/s40537-022-00592-5).
- Russia (2023) *Russian federal law N 149-ФЗ [On Information, Information Technologies, and Information Protection]*.
- Russia (2020) *Федеральный закон от 24.04.2020 № 123-ФЗ · Официальное опубликование правовых актов · Официальный интернет-портал правовой информации*. 2020.

<http://publication.pravo.gov.ru/Document/View/0001202004240030>

[Accessed: 2 August 2025].

- Ryabova, T. & Ryabov, O. (2011) The real man of politics in Russia (On gender discourse as a resource for the authority). *Social Sciences*. 42 (3), 58–71.
- Sætra, H.S. (2021) A Framework for Evaluating and Disclosing the ESG Related Impacts of AI with the SDGs. *Sustainability*. 13 (15), 8503.
- Sætra, H.S. (2020) A shallow defence of a technocracy of artificial intelligence: Examining the political harms of algorithmic governance in the domain of government. *Technology in Society*. 62, 101283.
- Sætra, H.S. & Selinger, E. (2024) Technological Remedies for Social Problems: Defining and Demarcating Techno-Fixes and Techno-Solutionism. *Science and Engineering Ethics*. 30 (6), 60. doi:[10.1007/s11948-024-00524-x](https://doi.org/10.1007/s11948-024-00524-x).
- Sætra, H.S. & Selinger, E. (2023) The Siren song of technological remedies for social problems: defining, demarcating, and evaluating techno-fixes and techno-solutionism. *Demarcating, and Evaluating Techno-Fixes and Techno-Solutionism* (September 19, 2023). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4576687.
- Sakaki, T., Okazaki, M. & Matsuo, Y. (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: *Proceedings of the 19th international conference on World wide web*. 2010 pp. 851–860.
- Sakieh, Y. (2023) Shaping climate change discourse: the nexus between political media landscape and recommendation systems in social networks. *Social Network Analysis and Mining*. 14 (1), 10. doi:[10.1007/s13278-023-01175-9](https://doi.org/10.1007/s13278-023-01175-9).
- Salvi, M., Acharya, U.R., Molinari, F. & Meiburger, K.M. (2021) The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Computers in Biology and Medicine*. 128, 104129.
- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. & Aroyo, L.M. (2021) “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 6 May 2021 Yokohama Japan, ACM. pp. 1–15. doi:[10.1145/3411764.3445518](https://doi.org/10.1145/3411764.3445518).

- Sánchez-Núñez, P., Cobo, M.J., Heras-Pedrosa, C.D.L., Peláez, J.I. & Herrera-Viedma, E. (2020) Opinion Mining, Sentiment Analysis and Emotion Understanding in Advertising: A Bibliometric Analysis. *IEEE Access*. 8, 134563–134576. doi:[10.1109/ACCESS.2020.3009482](https://doi.org/10.1109/ACCESS.2020.3009482).
- Saveliev, A. & Zhurenkov, D. (2021) Artificial intelligence and social responsibility: the case of the artificial intelligence strategies in the United States, Russia, and China. *Kybernetes*. 50 (3), 656–675.
- Saxe, A.M., McClelland, J.L. & Ganguli, S. (2013) Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.
- Scassa, T. (2018) Data ownership. *CIGI papers*. (187). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3251542.
- Schaar, P. (2010) Privacy by Design. *Identity in the Information Society*. 3 (2), 267–274. doi:[10.1007/s12394-010-0055-x](https://doi.org/10.1007/s12394-010-0055-x).
- Schafer, J.B., Frankowski, D., Herlocker, J. & Sen, S. (2007) Collaborative filtering recommender systems. In: *The adaptive web*. Springer. pp. 291–324.
- Schein, A.I., Popescul, A., Ungar, L.H. & Pennock, D.M. (2002) Methods and metrics for cold-start recommendations. In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 2002 pp. 253–260.
- Schelenz, L. (2021) Diversity-aware Recommendations for Social Justice? Exploring User Diversity and Fairness in Recommender Systems. In: *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 21 June 2021 Utrecht Netherlands, ACM. pp. 404–410. doi:[10.1145/3450614.3463293](https://doi.org/10.1145/3450614.3463293).
- Scherer, M.U. (2015) Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harv. JL & Tech*. 29, 353.
- Schiff, D.S. (2024) Framing contestation and public influence on policymakers: evidence from US artificial intelligence policy discourse. *Policy and Society*. 43 (3), 255–288.
- Schlenkrich, O. (2019) Identifying profiles of democracies: A cluster analysis based on the democracy matrix dataset from 1900 to 2017. *Politics and Governance*. 7 (4), 315–330.
- Schlozman, K.L., Verba, S. & Brady, H.E. (1999) *Civic participation and the equality problem*. Brookings Institution Press Washington, DC.

- Schmitt, J.B., Rieger, D., Rutkowski, O. & Ernst, J. (2018a) Counter-messages as prevention or promotion of extremism?! the potential role of youtube: Recommendation algorithms. *Journal of communication*. 68 (4), 780–808.
- Schmitt, J.B., Rieger, D., Rutkowski, O. & Ernst, J. (2018b) Counter-messages as prevention or promotion of extremism?! The potential role of YouTube: Recommendation algorithms. *Journal of communication*. 68 (4), 780–808.
- Schöning, J. & Kruse, N. (2025) *Compliance of AI Systems*. doi:[10.48550/arXiv.2503.05571](https://doi.org/10.48550/arXiv.2503.05571).
- Schröder, G., Thiele, M. & Lehner, W. (2011) Setting goals and choosing metrics for recommender system evaluations. In: *UCERST12 workshop at the 5th ACM conference on recommender systems, Chicago, USA*. 2011 p. 53.
- Schuett, J. (2023) Risk management in the artificial intelligence act. *European Journal of Risk Regulation*. 1–19.
- Schuman, E. (2024) *Dutch appellate court rules against Oracle and Salesforce in a GDPR-related cookie case*. 2024. CSO Online. <https://www.csoononline.com/article/2167475/dutch-appellate-court-rules-against-oracle-and-salesforce-in-a-gdpr-related-cookie-case.html> [Accessed: 12 July 2025].
- Schwartz, S. (2006) A theory of cultural value orientations: Explication and applications. *Comparative sociology*. 5 (2–3), 137–182.
- Schwartz, S.H. (1992) Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In: *Advances in experimental social psychology*. Elsevier. pp. 1–65.
- Sclafani, J. (2017) *Talking Donald Trump: A sociolinguistic study of style, metadiscourse, and political identity*. Routledge. <https://www.taylorfrancis.com/books/mono/10.4324/9781315276885/talking-donald-trump-jennifer-sclafani>.
- Scott, C. (2018) Accountability in the regulatory state. In: *Administrative law*. Routledge. pp. 397–419. <https://www.taylorfrancis.com/chapters/edit/10.4324/9781315183770-17/accountability-regulatory-state-colin-scott>.

- Scott, S.L. (2010) A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*. 26 (6), 639–658. doi:[10.1002/asmb.874](https://doi.org/10.1002/asmb.874).
- Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F. & Dennison, D. (2015) Hidden technical debt in machine learning systems. *Advances in neural information processing systems*. 28.
- Seabright, P. (1996) Accountability and decentralisation in government: An incomplete contracts model. *The Domain of the State*. 40 (1), 61–89. doi:[10.1016/0014-2921\(95\)00055-0](https://doi.org/10.1016/0014-2921(95)00055-0).
- Sebastian, A.M. & Peter, D. (2022) Artificial Intelligence in Cancer Research: Trends, Challenges and Future Directions. *Life*. 12 (12). doi:[10.3390/life12121991](https://doi.org/10.3390/life12121991).
- Sedlmair, M., Meyer, M. & Munzner, T. (2012) Design study methodology: Reflections from the trenches and the stacks. *IEEE transactions on visualization and computer graphics*. 18 (12), 2431–2440.
- Selbst, A.D. (2017) Disparate impact in big data policing. *Ga. L. Rev.* 52, 109.
- Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. & Vertesi, J. (2019) Fairness and Abstraction in Sociotechnical Systems. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29 January 2019 Atlanta GA USA, ACM. pp. 59–68. doi:[10.1145/3287560.3287598](https://doi.org/10.1145/3287560.3287598).
- Semenova, E., Perevoshchikova, E., Ivanov, A. & Erofeev, M. (2019) *Fairness meets machine learning: searching for a better balance*. <https://wp.hse.ru/data/2019/12/12/1524266321/93LAW2019.pdf>.
- Semykina, K. (2021) *The Crimean Conflict on Russian and Ukrainian TV: A Discourse-Theoretical Approach*. https://discourseanalysis.net/sites/default/files/2021-11/Semykina_2021_DNCWPS_6.pdf.
- Sen, A. (2017) Democracy as a universal value. In: *Applied Ethics*. Routledge. pp. 107–117.
- Setiya, K. (2013) Epistemic agency: Some doubts. *Philosophical Issues*. 23, 179–198.

- Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J. & Sculley, D. (2017) No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*.
- Shastri, A. (2023) *Advances In AI, Yet Many Unresolved Issues*. 2023. Forbes. <https://www.forbes.com/sites/aranshastri/2023/12/10/advances-in-ai-yet-many-unresolved-issues/> [Accessed: 19 July 2025].
- Shen, A., Han, X., Cohn, T., Baldwin, T. & Frermann, L. (2022) *Optimising Equal Opportunity Fairness in Model Training*. doi:[10.48550/arXiv.2205.02393](https://doi.org/10.48550/arXiv.2205.02393).
- Siegmann, C. & Anderljung, M. (2022) *The Brussels Effect and Artificial Intelligence: How EU regulation will impact the global AI market*. doi:[10.48550/arXiv.2208.12645](https://doi.org/10.48550/arXiv.2208.12645).
- Simmhan, Y.L., Plale, B. & Gannon, D. (2005) A survey of data provenance in e-science. *ACM Sigmod Record*. 34 (3), 31–36.
- Simonite, T. (2022) *Google Has a Plan to Stop Its New AI From Being Dirty and Rude* | WIRED. 2022. <https://www.wired.com/story/google-voice-assistant-future/> [Accessed: 7 July 2025].
- Singh, A. & Joachims, T. (2018) Fairness of Exposure in Rankings. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 19 July 2018 London United Kingdom, ACM. pp. 2219–2228. doi:[10.1145/3219819.3220088](https://doi.org/10.1145/3219819.3220088).
- Sinha, A., Gleich, D.F. & Ramani, K. (2016) Deconvolving feedback loops in recommender systems. *Advances in neural information processing systems*. 29. <https://proceedings.neurips.cc/paper/2016/hash/962e56a8a0b0420d87272a682bfd1e53-Abstract.html>.
- Sivetc, L. (2019) State regulation of online speech in Russia: The role of internet infrastructure owners. *International Journal of Law and Information Technology*. 27 (1), 28–49.
- Sivetc, L. & Wijermars, M. (2021) The Vulnerabilities of Trusted Notifier-Models in Russia: The Case of Netoscope. *Media and Communication*. 9 (4), 27–38. doi:[10.17645/mac.v9i4.4237](https://doi.org/10.17645/mac.v9i4.4237).
- Sky News Australia (2025) 'Going off talking points': 22 Democrats use same script in anti-Trump post. <https://www.youtube.com/watch?v=aokv0QHotrc>.

- Slachmuis, L. (2025) Can AI break the engagement trap? 14 August 2025. Tech and Social Cohesion.
https://techandsocialcohesion.substack.com/p/can-ai-break-the-engagement-trap?publication_id=2192522&utm_campaign=email-post-title&r=3a8s6d&utm_medium=email.
- Smith, R. (2007) An overview of the Tesseract OCR engine. In: *Ninth international conference on document analysis and recognition (ICDAR 2007)*. 2007 IEEE. pp. 629–633.
<https://ieeexplore.ieee.org/abstract/document/4376991/>.
- Smuha, N.A. (2021) From a ‘race to AI’ to a ‘race to AI regulation’: regulatory competition for artificial intelligence. *Law, Innovation and Technology*. 13 (1), 57–84. doi:[10.1080/17579961.2021.1898300](https://doi.org/10.1080/17579961.2021.1898300).
- Smyth, R. (2025) *EU firms disclosing more detail on use of artificial intelligence*. 23 November 2025. Irish Examiner.
<https://www.irishexaminer.com/business/companies/arid-41747604.html> [Accessed: 12 January 2026].
- Soares, A.M., Farhangmehr, M. & Shoham, A. (2007) Hofstede’s dimensions of culture in international marketing studies. *Journal of business research*. 60 (3), 277–284.
- Soldatov, A. & Borogan, I. (2015) Google-Books-ID: u7l4DgAAQBAJ. *The Red Web: The Struggle Between Russia’s Digital Dictators and the New Online Revolutionaries*. Hachette UK.
- Solsman, J.E. (2018) YouTube’s AI is the puppet master over most of what you watch. *CNET, January*. 10.
- Solum, L.B. (2020) Legal personhood for artificial intelligences. In: *Machine ethics and robot ethics*. Routledge. pp. 415–471.
<https://www.taylorfrancis.com/chapters/edit/10.4324/9781003074991-37/legal-personhood-artificial-intelligences-lawrence-solum>.
- Sonne, P. & Das, A. (2024) *Modi’s Russia Visit Showcases a Less Isolated Putin, Angering Ukraine - The New York Times*. 2024.
<https://www.nytimes.com/2024/07/09/world/europe/russia-india-moscow-putin.html> [Accessed: 19 July 2025].
- Sood, A., Forster, R.A., Archer, B.J. & Little, R.C. (2021) Neutronics Calculation Advances at Los Alamos: Manhattan Project to Monte Carlo. *Nuclear Technology*. 207 (sup1), S100–S133.
doi:[10.1080/00295450.2021.1956255](https://doi.org/10.1080/00295450.2021.1956255).

- Sousa, J., Santos, S., André, L. & Ferreira, J. (2024) Converging Affective Computing and Ethical Challenges: The Quest for Universal Access in Human-Machine Cooperation. In: M. Antona & C. Stephanidis (eds). *Universal Access in Human-Computer Interaction*. Lecture Notes in Computer Science. Cham, Springer Nature Switzerland. pp. 106–121. doi:[10.1007/978-3-031-60875-9_8](https://doi.org/10.1007/978-3-031-60875-9_8).
- Spiekermann, S., Acquisti, A., Böhme, R. & Hui, K.-L. (2015) The challenges of personal data markets and privacy. *Electronic Markets*. 25 (2), 161–167. doi:[10.1007/s12525-015-0191-0](https://doi.org/10.1007/s12525-015-0191-0).
- Srinivasan, R. & Chander, A. (2021) Biases in AI systems. *Communications of the ACM*. 64 (8), 44–49.
- Srivastava, Y., Murali, V. & Dubey, S.R. (2019) A performance evaluation of loss functions for deep face recognition. In: *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*. 2019 Springer. pp. 322–332.
- Stadnik, I. (2021a) Control by infrastructure: Political ambitions meet technical implementations in RuNet. *First Monday*. <https://firstmonday.org/ojs/index.php/fm/article/view/11693>.
- Stadnik, I. (2021b) Russia: An independent and sovereign internet? In: *Power and authority in internet governance*. Routledge. pp. 147–167. <https://www.taylorfrancis.com/chapters/edit/10.4324/9781003008309-10/russia-ilona-stadnik>.
- Star, S.L. & Griesemer, J.R. (1989) Institutional Ecology, ‘Translations’ and Boundary Objects: Amateurs and Professionals in Berkeley’s Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science*. 19 (3), 387–420. doi:[10.1177/030631289019003001](https://doi.org/10.1177/030631289019003001).
- Steck, H. (2011) Item popularity and recommendation accuracy. In: *Proceedings of the fifth ACM conference on Recommender systems*. 23 October 2011 Chicago Illinois USA, ACM. pp. 125–132. doi:[10.1145/2043932.2043957](https://doi.org/10.1145/2043932.2043957).
- Stepanova, E. (2015) ‘The Spiritual and Moral Foundation of Civilization in Every Nation for Thousands of Years’: The Traditional Values Discourse in Russia. *Politics, Religion & Ideology*. 16 (2–3), 119–136. doi:[10.1080/21567689.2015.1068167](https://doi.org/10.1080/21567689.2015.1068167).
- Stray, J. (2021) *Designing Recommender Systems to Depolarize*. doi:[10.48550/arXiv.2107.04953](https://doi.org/10.48550/arXiv.2107.04953).

- Stromer-Galley, J. (2003) Diversity of political conversation on the Internet: Users' perspectives. *Journal of Computer-Mediated Communication*. 8 (3), JCMC836.
- Sullivan, C. (2019) EU GDPR or APEC CBPR? A comparative analysis of the approach of the EU and APEC to cross border data transfers and protection of personal data in the IoT era. *Computer Law & Security Review*. 35 (4), 380–397. doi:[10.1016/j.clsr.2019.05.004](https://doi.org/10.1016/j.clsr.2019.05.004).
- Sullivan, F. (2014) *The utilization of sock puppets in cyber intelligence operation*. Master's Thesis. Utica College.
<https://search.proquest.com/openview/fa9af073720e144680b04a7aa59ea851/1?pg-origsite=gscholar&cbl=18750>.
- Sun, J. (2024) *Council Post: Apple Intelligence: A Double-Edged Sword For The Commerce Landscape*. 2024. Forbes.
<https://www.forbes.com/sites/forbestechcouncil/2024/07/09/apple-intelligence-a-double-edged-sword-for-the-commerce-landscape/>
[Accessed: 9 July 2025].
- Svantesson, D. (2022) The European Union Artificial Intelligence Act: Potential implications for Australia. *Alternative Law Journal*. 47 (1), 4–9. doi:[10.1177/1037969X211052339](https://doi.org/10.1177/1037969X211052339).
- Świechowski, M., Godlewski, K., Sawicki, B. & Mańdziuk, J. (2023) Monte Carlo Tree Search: a review of recent modifications and applications. *Artificial Intelligence Review*. 56 (3), 2497–2562. doi:[10.1007/s10462-022-10228-y](https://doi.org/10.1007/s10462-022-10228-y).
- Sylwester, K. & Purver, M. (2015) Twitter language use reflects psychological differences between democrats and republicans. *PloS one*. 10 (9), e0137422.
- Symeonidis, P. (2008) Content-based dimensionality reduction for recommender systems. In: *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation eV, Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*. 2008 Springer. pp. 619–626.
- Taeihagh, A. (2021) Governance of artificial intelligence. *Policy and society*. 40 (2), 137–157.
- Tandon, R., Singh, W., Clover, D.E. & Hall, B.L. (2016) *Knowledge democracy and excellence in engagement*.

- Taskesen, B., Blanchet, J., Kuhn, D. & Nguyen, V.A. (2021) A Statistical Test for Probabilistic Fairness. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 3 March 2021 Virtual Event Canada, ACM. pp. 648–665. doi:[10.1145/3442188.3445927](https://doi.org/10.1145/3442188.3445927).
- Teper, Y. (2016) Official Russian identity discourse in light of the annexation of Crimea: national or imperial? *Post-Soviet Affairs*. 32 (4), 378–396. doi:[10.1080/1060586X.2015.1076959](https://doi.org/10.1080/1060586X.2015.1076959).
- Terren, L. & Borge, R. (2021) *Echo chambers on social media: A systematic review of the literature*. <https://openaccess.uoc.edu/items/4703db77-37f3-488a-8e90-8d680bc2d497>.
- The Associated Press (2024) *Japan's Kishida unveils a framework for global regulation of generative AI*. 2 May 2024. AP News. <https://apnews.com/article/oecd-ai-japan-kishida-artificial-intelligence-023ac08e04db5a2109cf35f8b8c9b102> [Accessed: 19 July 2025].
- The Boston Globe (2024) *Lawsuits challenge how those internet cookies crumble - The Boston Globe*. 2024. BostonGlobe.com. <https://www.bostonglobe.com/2024/04/03/opinion/data-collection-lawsuit-new-england-baptist-beth-israel-hospital/> [Accessed: 12 July 2025].
- The New York Times Editorial (2007) Opinion | Exit, Russian Democracy. *The New York Times*. 27 November. <https://www.nytimes.com/2007/11/27/opinion/27tue2.html>.
- The White House, T.W. (2023) *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. 30 October 2023. The White House. <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/> [Accessed: 9 June 2025].
- Thierer, A.D., Castillo O'Sullivan, A. & Russell, R. (2017) Artificial intelligence and public policy. *Mercatus Research Paper*.
- Thomas, D.G., Ellen (2024) *In an AI arms race, data centers are going nuclear*. 2024. Business Insider. <https://www.businessinsider.com/data-centers-amazon-ai-nuclear-power-grid-2024-5> [Accessed: 11 July 2025].

- Thompson, J.J., Wilby, R.L., Matthews, T. & Murphy, C. (2022) The utility of Google Trends as a tool for evaluating flooding in data-scarce places. *Area*. 54 (2), 203–212. doi:[10.1111/area.12719](https://doi.org/10.1111/area.12719).
- Thorat, P.B., Goudar, R.M. & Barve, S. (2015) Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*. 110 (4), 31–36.
- Tian, K. (2020) *Cross-cultural recommender system use and design*. PhD Thesis. University of Illinois at Urbana-Champaign. <https://www.ideals.illinois.edu/items/115770>.
- Toderici, G., Aradhye, H., Pasca, M., Sbaiz, L. & Yagnik, J. (2010) Finding meaning on youtube: Tag recommendation and category discovery. In: *2010 IEEE computer society conference on computer vision and pattern recognition*. 2010 IEEE. pp. 3447–3454. <https://ieeexplore.ieee.org/abstract/document/5539985/>.
- Tom, E., Aurum, A. & Vidgen, R. (2013) An exploration of technical debt. *Journal of Systems and Software*. 86 (6), 1498–1516.
- Tomar, D. & Agarwal, S. (2014) A survey on pre-processing and post-processing techniques in data mining. *International Journal of Database Theory & Application*. 7 (4), 99–128.
- Tomaschek, A., Long, M. & Hodge, R. (2024) *These VPNs Can Help Improve Your Online Privacy*. 2024. CNET. <https://www.cnet.com/tech/services-and-software/best-vpn/> [Accessed: 19 July 2025].
- Toyama, K. (2011) Technology as amplifier in international development. In: *Proceedings of the 2011 iConference*. 8 February 2011 Seattle Washington USA, ACM. pp. 75–82. doi:[10.1145/1940761.1940772](https://doi.org/10.1145/1940761.1940772).
- Treat, R.M., Hsiao, S.K., Ismail, A. & Javan, R. (2024) The US Government's Latest Presidential Executive Order on Artificial Intelligence: Potential Implications in Radiology. *Journal of the American College of Radiology*. <https://www.sciencedirect.com/science/article/pii/S1546144024003557/pdf?md5=b520a409da40eeed026584420e79dfa2&pid=1-s2.0-S1546144024003557-main.pdf>.
- Tsai, C.I., Klayman, J. & Hastie, R. (2008) Effects of amount of information on judgment accuracy and confidence. *Organizational*

Behavior and Human Decision Processes. 107 (2), 97–105.

doi:[10.1016/j.obhdp.2008.01.005](https://doi.org/10.1016/j.obhdp.2008.01.005).

- Tsygankov, A.P. (2014) Google-Books-ID: HnvOBAAAQBAJ. *The Strong State in Russia: Development and Crisis*. Oxford University Press.
- Tsygankov, A.P. & Parker, D. (2015) The securitization of democracy: Freedom House ratings of Russia. *European Security*. 24 (1), 77–100. doi:[10.1080/09662839.2014.968775](https://doi.org/10.1080/09662839.2014.968775).
- Tufekci, Z. (2017) *Twitter and tear gas: The power and fragility of networked protest*. Yale University Press.
- Tufekci, Z. (2018) YouTube, the great radicalizer. *The New York Times*. 10 (3), 2018.
- Tung, R.L. & Verbeke, A. (2010) Beyond Hofstede and GLOBE: Improving the quality of cross-cultural research. *Journal of International Business Studies*. 41 (8), 1259–1274. doi:[10.1057/jibs.2010.41](https://doi.org/10.1057/jibs.2010.41).
- Ulgen, O. (2020) User Rights and Adaptive A/IS – From Passive Interaction to Real Empowerment. In: R.A. Sottilare & J. Schwarz (eds). *Adaptive Instructional Systems*. Lecture Notes in Computer Science. Cham, Springer International Publishing. pp. 205–217. doi:[10.1007/978-3-030-50788-6_15](https://doi.org/10.1007/978-3-030-50788-6_15).
- Unit, E.I. (2024) *Democracy index 2023: Age of conflict*. <https://policycommons.net/artifacts/11334740/democracy-index-2023-final-report/12223667/>.
- US (2022) *Blueprint for an AI Bill of Rights* | OSTP. 2022. The White House. <https://www.whitehouse.gov/ostp/ai-bill-of-rights/> [Accessed: 1 February 2025].
- Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K.-L.A., Elkhatab, Y., Hussain, A. & Al-Fuqaha, A. (2019) Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE access*. 7, 65579–65615.
- USCode (2024) *15 USC 9401: Definitions*. 2024. [https://uscode.house.gov/view.xhtml?req=\(title:15%20section:9401%20edition:prelim\)](https://uscode.house.gov/view.xhtml?req=(title:15%20section:9401%20edition:prelim)) [Accessed: 19 July 2025].
- Vainionpää, F., Väyrynen, K., Lanamäki, A. & Parmiggiani, E. (2024) Practices of anticipation: How public sector organizations anticipate

artificial intelligence and its regulation. *ECIS 2024 Proceedings*.
<https://oulurepo.oulu.fi/handle/10024/49294>.

- Vakkuri, V., Kemell, K.-K., Jantunen, M., Halme, E. & Abrahamsson, P. (2021) ECCOLA—A method for implementing ethically aligned AI systems. *Journal of Systems and Software*. 182, 111067.
- Valstar, M. (2024) *What does the EU AI Act mean for Affective Computing and Emotion AI?* 2024.
<https://www.linkedin.com/pulse/what-does-eu-ai-act-mean-affective-computing-emotion-michel-valstar-pblue> [Accessed: 19 July 2025].
- Van Caneghem, T. & Van Campenhout, G. (2012) Quantity and quality of information and SME financial structure. *Small Business Economics*. 39 (2), 341–358.
- Van der Aalst, W.M., Rubin, V., Verbeek, H.M.W., van Dongen, B.F., Kindler, E. & Günther, C.W. (2010) Process mining: a two-step approach to balance between underfitting and overfitting. *Software & Systems Modeling*. 9, 87–111.
- Van Eijk, N., Helberger, N., Kool, L., van der Plas, A. & van der Sloot, B. (2012) Online tracking: questioning the power of informed consent. *info*. 14 (5), 57–73.
- Van Maanen, G. (2022) AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics. *Digital Society*. 1 (2), 9. doi:[10.1007/s44206-022-00013-3](https://doi.org/10.1007/s44206-022-00013-3).
- Varshney, L.R. (2020) *Respect for Human Autonomy in Recommender Systems*. doi:[10.48550/arXiv.2009.02603](https://doi.org/10.48550/arXiv.2009.02603).
- Vartanova, E. (2019) Russian media: a call for theorising the economic change. *Russian Journal of Communication*. 11 (1), 22–36.
doi:[10.1080/19409419.2019.1572531](https://doi.org/10.1080/19409419.2019.1572531).
- Vartanova, E., Gladkova, A., Lapin, D., Samorodova, E. & Vikhrova, O. (2021) Theorizing Russian model of the digital divide. *World of Media. Journal of Russian Media and Journalism Studies*. (1), 5–40.
- Veale, M., Binns, R. & Ausloos, J. (2018) When data protection by design and data subject rights clash. *International Data Privacy Law*. 8 (2), 105–123.
- Veale, M. & Zuiderveen Borgesius, F. (2021) Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the

unclear elements of the proposed approach. *Computer Law Review International*. 22 (4), 97–112. doi:[10.9785/cri-2021-220402](https://doi.org/10.9785/cri-2021-220402).

- Velez-Estevez, A., Perez, I.J., García-Sánchez, P., Moral-Muñoz, J.A. & Cobo, M.J. (2023) New trends in bibliometric APIs: A comparative analysis. *Information Processing & Management*. 60 (4), 103385.
- Venkatesan, K.G.S. (2013) Comparison of CDMA & GSM Mobile Technology. *Middle-East Journal of Scientific Research*. 13 (12), 1590–1594.
- Vente, T. (2023) Advancing Automation of Design Decisions in Recommender System Pipelines. In: *Proceedings of the 17th ACM Conference on Recommender Systems*. 14 September 2023 Singapore Singapore, ACM. pp. 1355–1360. doi:[10.1145/3604915.3608886](https://doi.org/10.1145/3604915.3608886).
- Vente, T., Ekstrand, M. & Beel, J. (2023) Introducing LensKit-Auto, an Experimental Automated Recommender System (AutoRecSys) Toolkit. In: *Proceedings of the 17th ACM Conference on Recommender Systems*. 14 September 2023 Singapore Singapore, ACM. pp. 1212–1216. doi:[10.1145/3604915.3610656](https://doi.org/10.1145/3604915.3610656).
- Viljoen, S. (2021) A relational theory of data governance. *The Yale Law Journal*. 573–654.
- Vincent, J. (2023) *Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content*. 17 January 2023. The Verge. <https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit> [Accessed: 21 July 2025].
- VisitUkraine (2023) *The Ukrainian language is becoming more prevalent on social media: more than 90% of Instagram posts in 2023 were created in Ukrainian*. 2023. <https://visitukraine.today/pl/blog/2834/the-ukrainian-language-is-becoming-more-prevalent-on-social-media-more-than-90-of-instagram-posts-in-2023-were-created-in-ukrainian> [Accessed: 28 June 2025].
- Viswanath, K. & Finnegan Jr, J.R. (1996) The knowledge gap hypothesis: Twenty-five years later. *Annals of the International Communication Association*. 19 (1), 187–228.
- Wachter, S., Mittelstadt, B. & Russell, C. (2021) Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review*. 41, 105567.

- Walker, D.L., Geertsema, M. & Barnett, B. (2009) Inverting the inverted pyramid: A conversation about the use of feminist theories to teach journalism. *Feminist Teacher*. 19 (3), 177–194.
- Wang, P. (2019) On defining artificial intelligence. *Journal of Artificial General Intelligence*. 10 (2), 1–37.
- Wang, Y. & Ruhe, G. (2007) The cognitive process of decision making. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*. 1 (2), 73–85.
- Wankhade, M., Rao, A.C.S. & Kulkarni, C. (2022) A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*. 55 (7), 5731–5780. doi:[10.1007/s10462-022-10144-1](https://doi.org/10.1007/s10462-022-10144-1).
- Ward Jr, J.H. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*. 58 (301), 236–244.
- Waytz, A., Iyer, R., Young, L., Haidt, J. & Graham, J. (2019) Ideological differences in the expanse of the moral circle. *Nature Communications*. 10 (1), 4389.
- Weber, M. (2019) *Economy and society: A new translation*. Harvard University Press.
- Weber, M. (1978) *Economy and society: An outline of interpretive sociology*. University of California press.
- Webster, F. (2002) The information society revisited. *Handbook of new media: Social shaping and social consequences of ICTs*. 22–33.
- Webster, F. (2014) *Theories of the information society*. Routledge.
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., et al. (2021) *Ethical and social risks of harm from Language Models*. doi:[10.48550/arXiv.2112.04359](https://doi.org/10.48550/arXiv.2112.04359).
- Werbos, P. (1974) Beyond regression:" new tools for prediction and analysis in the behavioral sciences. *Ph. D. dissertation, Harvard University*.
- Werbos, P.J. (1990) Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*. 78 (10), 1550–1560.
- Whimster, S. & Lash, S. (2014) *Max Weber, rationality and modernity*. Routledge.

<https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.4324/9781315823935&type=googlepdf>.

- Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S.M., Richardson, R., Schultz, J. & Schwartz, O. (2018) *AI now report 2018*. AI Now Institute at New York University New York.
https://www.stc.org/roundtable/wp-content/uploads/sites/34/2019/06/AI_Now_2018_Report.pdf.
- Wiggers, K. (2023) Researchers discover a way to make ChatGPT consistently toxic. 12 April 2023. TechCrunch.
<https://techcrunch.com/2023/04/12/researchers-discover-a-way-to-make-chatgpt-consistently-toxic/>.
- Wilkinson, C. (2014) Putting “Traditional Values” Into Practice: The Rise and Contestation of Anti-Homopropaganda Laws in Russia. *Journal of Human Rights*. 13 (3), 363–379. doi:[10.1080/14754835.2014.919218](https://doi.org/10.1080/14754835.2014.919218).
- Wilson, T.P. (2017) Normative and interpretive paradigms in sociology. In: *Everyday life*. Routledge. pp. 57–79.
<https://www.taylorfrancis.com/chapters/edit/10.4324/9781351327329-3/normative-interpretive-paradigms-sociology-thomas-wilson>.
- Winner, L. (2020) *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press.
<https://press.uchicago.edu/ucp/books/book/chicago/W/bo49911830.html>
!
- Wirtschafter, E.K. (2006) Russian legal culture and the rule of law. *Kritika: Explorations in Russian and Eurasian History*. 7 (1), 61–70.
- Wischmeyer, T. & Rademacher, T. (2020) *Regulating Artificial Intelligence*. Cham, Springer International Publishing. doi:[10.1007/978-3-030-32361-5](https://doi.org/10.1007/978-3-030-32361-5).
- Woch, M., Zieja, M. & Tomaszewska, J. (2017) Analysis of the Time between Failures of Aircrafts. In: *2017 2nd International Conference on System Reliability and Safety (ICSRS)*. 2017 IEEE. pp. 112–118.
<https://ieeexplore.ieee.org/abstract/document/8272805/>.
- Wojcieszak, M.E. & Mutz, D.C. (2009) Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of communication*. 59 (1), 40–56.

- Wolff, A., Gooch, D., Montaner, J.J.C., Rashid, U. & Kortuem, G. (2016) Creating an understanding of data literacy for a data-driven society. *The Journal of Community Informatics*. 12 (3).
- Wright, J. (2024) The Development of AI Ethics in Japan: Ethics-washing Society 5.0? *East Asian Science, Technology and Society: An International Journal*. 18 (2), 117–134.
doi:[10.1080/18752160.2023.2275987](https://doi.org/10.1080/18752160.2023.2275987).
- Wu, C.-Y., Ahmed, A., Beutel, A., Smola, A.J. & Jing, H. (2017) Recurrent Recommender Networks. In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 2 February 2017 Cambridge United Kingdom, ACM. pp. 495–503.
doi:[10.1145/3018661.3018689](https://doi.org/10.1145/3018661.3018689).
- Wu, H. & Zhou, J. (2021) Privacy leakage of SIFT features via deep generative model based image reconstruction. *IEEE Transactions on Information Forensics and Security*. 16, 2973–2985.
- Wu, M. (2006) Hofstede's cultural dimensions 30 years later: A study of Taiwan and the United States. *Intercultural communication studies*. 15 (1), 33.
- Wu, M., Liu, W., Wang, Y. & Yao, M. (2024) *Negotiating the Shared Agency between Humans & AI in the Recommender System*.
doi:[10.48550/arXiv.2403.15919](https://doi.org/10.48550/arXiv.2403.15919).
- Wu, Y., Cao, J. & Xu, G. (2024) Fairness in Recommender Systems: Evaluation Approaches and Assurance Strategies. *ACM Transactions on Knowledge Discovery from Data*. 18 (1), 1–37. doi:[10.1145/3604558](https://doi.org/10.1145/3604558).
- Xu, C., Zhao, P., Liu, Y., Xu, J., S.Sheng, V.S.S., Cui, Z., Zhou, X. & Xiong, H. (2019) Recurrent Convolutional Neural Network for Sequential Recommendation. In: *The World Wide Web Conference*. 13 May 2019 San Francisco CA USA, ACM. pp. 3398–3404.
doi:[10.1145/3308558.3313408](https://doi.org/10.1145/3308558.3313408).
- Xu, H., Liu, X., Li, Y., Jain, A. & Tang, J. (2021) To be robust or to be fair: Towards fairness in adversarial training. In: *International conference on machine learning*. 2021 PMLR. pp. 11492–11501.
<http://proceedings.mlr.press/v139/xu21b.html>.
- Yamaguchi, E.P., Hurst, H.M. & Spielman, I.B. (2023) Feedback cooled Bose-Einstein condensation: near and far from equilibrium. *Physical Review A*. 107 (6), 063306. doi:[10.1103/PhysRevA.107.063306](https://doi.org/10.1103/PhysRevA.107.063306).

- Yang, S. & Berdine, G. (2017) The receiver operating characteristic (ROC) curve. *The Southwest Respiratory and Critical Care Chronicles*. 5 (19), 34–36.
- Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W.L. & Leskovec, J. (2018) Graph convolutional neural networks for web-scale recommender systems. In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018 pp. 974–983.
- Yohe, K.A. (2023) *Towards Global, Socio-Economic, and Culturally Aware Recommender Systems*. doi:[10.48550/arXiv.2312.05805](https://doi.org/10.48550/arXiv.2312.05805).
- Younis, R., Ahmad, A. & Abu Al-Haija, Q. (2022) Explaining intrusion detection-based convolutional neural networks using shapley additive explanations (shap). *Big Data and Cognitive Computing*. 6 (4), 126.
- YouTube Team (2019) *Continuing our work to improve recommendations on YouTube*. 2019. [blog.youtube](https://blog.youtube/news-and-events/continuing-our-work-to-improve/).
<https://blog.youtube/news-and-events/continuing-our-work-to-improve/> [Accessed: 3 June 2025].
- Ytreberg, E. (2001) Moving out of the Inverted Pyramid: narratives and descriptions in television news. *Journalism Studies*. 2 (3), 357–371. doi:[10.1080/14616700118194](https://doi.org/10.1080/14616700118194).
- Yu, X., Haroon, M., Menchen-Trevino, E. & Wojcieszak, M. (2024) Nudging recommendation algorithms increases news consumption and diversity on YouTube. *PNAS nexus*. 3 (12), pgae518.
- Yuan, S., Wang, J. & Zhao, X. (2013) Real-time bidding for online advertising: measurement and analysis. In: *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*. ADKDD '13. 11 August 2013 New York, NY, USA, Association for Computing Machinery. pp. 1–8. doi:[10.1145/2501040.2501980](https://doi.org/10.1145/2501040.2501980).
- Yudkowsky, E. (2016) The AI alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*. 4, 1.
- Zangerle, E. & Bauer, C. (2022) Evaluating Recommender Systems: Survey and Framework. *ACM Computing Surveys*. 55 (8), 1–38.
- Zhang, J. & Wedel, M. (2009) The Effectiveness of Customized Promotions in Online and Offline Stores. *Journal of Marketing Research*. 46 (2), 190–206. doi:[10.1509/jmkr.46.2.190](https://doi.org/10.1509/jmkr.46.2.190).

- Zhang, S., Yao, L., Sun, A. & Tay, Y. (2019) Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*. 52 (1), 1–38.
- Zhang, W., Qin, J., Guo, W., Tang, R. & He, X. (2021) *Deep Learning for Click-Through Rate Estimation*. doi:[10.48550/arXiv.2104.10584](https://doi.org/10.48550/arXiv.2104.10584).
- Zhang, W., Yuan, S. & Wang, J. (2014) Optimal real-time bidding for display advertising. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '14. 24 August 2014 New York, NY, USA, Association for Computing Machinery. pp. 1077–1086. doi:[10.1145/2623330.2623633](https://doi.org/10.1145/2623330.2623633).
- Zhang, Y., Shah, D., Pevehouse, J. & Valenzuela, S. (2023) Reactive and Asymmetric Communication Flows: Social Media Discourse and Partisan News Framing in the Wake of Mass Shootings. *The International Journal of Press/Politics*. 28 (4), 837–861. doi:[10.1177/19401612211072793](https://doi.org/10.1177/19401612211072793).
- Zhao, W., Alwidian, S. & Mahmoud, Q.H. (2022) Adversarial training methods for deep learning: A systematic review. *Algorithms*. 15 (8), 283.
- Zhao, Y., Wang, Y., Liu, Y., Cheng, X., Aggarwal, C.C. & Derr, T. (2025) Fairness and Diversity in Recommender Systems: A Survey. *ACM Transactions on Intelligent Systems and Technology*. 16 (1), 1–28. doi:[10.1145/3664928](https://doi.org/10.1145/3664928).
- Zhao, Z., Hong, L., Wei, L., Chen, J., Nath, A., Andrews, S., Kumthekar, A., Sathiamoorthy, M., Yi, X. & Chi, E. (2019) Recommending what video to watch next: a multitask ranking system. In: *Proceedings of the 13th ACM Conference on Recommender Systems*. 10 September 2019 Copenhagen Denmark, ACM. pp. 43–51. doi:[10.1145/3298689.3346997](https://doi.org/10.1145/3298689.3346997).
- Zhdaneev, O.V. (2022) Technological sovereignty of the Russian Federation fuel and energy complex. *Записки Горного института*. 258, 1049–1066.
- Zhou, K., Meitus, A.A., Chase, M., Wang, G., Mykland, A., Howell, W. & Tan, C. (2024) Quantifying the Uniqueness and Divisiveness of Presidential Discourse. *PNAS Nexus*. 3 (10), pgae431. doi:[10.1093/pnasnexus/pgae431](https://doi.org/10.1093/pnasnexus/pgae431).

- Zhou, R., Khemmarat, S. & Gao, L. (2010) The impact of YouTube recommendation system on video views. In: *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 2010 pp. 404–410.
- Zhou, X.S. & Huang, T.S. (2003) Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems*. 8, 536–544.
- Ziegler, C.-N., McNee, S.M., Konstan, J.A. & Lausen, G. (2005) Improving recommendation lists through topic diversification. In: *Proceedings of the 14th international conference on World Wide Web*. 2005 pp. 22–32.
- Ziegler, W. (2020) A Landscape Analysis of Standardisation in the Field of Artificial Intelligence. *Journal of ICT Standardization*. 8 (2), 151–184.
- Zins, A.H., Bauernfeind, U., Del Missier, F., Venturini, A., Rumetshofer, H. & Frew, A.J. (2004) An experimental usability test for different destination recommender systems. In: *ENTER*. 2004 Citeseer. pp. 228–238.
<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f3ad9f54639e03a81a9d66223b79a59b86a47791>.
- Zou, J. & Schiebinger, L. (2018) *AI can be sexist and racist—it's time to make it fair*. Nature Publishing Group.
- Zuboff, S. (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, edn. *PublicAffairs, New York*. <https://www.academia.edu/download/61529191/GoodReads-The-Age-Of-Surveillance-Capitalism-The-Fight-For-A-Human-Future-At-The-New-Frontier-Of-Po20191216-80044-105m52n.pdf>.
- Zuiderveen Borgesius, F.J. (2020) Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights*. 24 (10), 1572–1593.
doi:[10.1080/13642987.2020.1743976](https://doi.org/10.1080/13642987.2020.1743976).
- Акимова, В.А., Бережной, В.В., Бородина, Е.М., Бутковский, Ю.В., Емелина, А.В., Ивашина, А.А., Карташова, Г.А., Конева, Е.В., Кроличенко, А.А., Мельников, Э.А., Старостин, В.С. & Черенкова, Е.А. (2018) *Использование Искусственного Интеллекта И Цифровой Аналитики Для Совершенствования Учебного Процесса И Его Инфраструктуры В Вузе*. Государственный университет управления. <https://elibrary.ru/item.asp?id=36783844>.

- Иванов, Д.В., Корженяк, А.М. & Lapikhina, E.S. (2021) Lethal Autonomous Weapons Systems and International Law. *Moscow Journal of International Law*. (3), 6–19. doi:[10.24833/0869-0049-2021-3-6-19](https://doi.org/10.24833/0869-0049-2021-3-6-19).
- Килячков, А.А., Чалдаева, Л.А., Королев, Д.А. & Байер, А.В. (2021) Использование искусственного интеллекта для выявления признаков мошеннических действий в ходе выборов. *Власть*. (5), 128–132.
- Колесников, Р. & Ваганова, В. (2022) «Государство просто де-факто забирает «Яндекс»: Путин нашел «идеальное решение» для Кудрина. 25 November 2022. БИЗНЕС Online. <https://www.business-gazeta.ru/article/573369> [Accessed: 18 April 2025].
- Панфилов, А.В., Дерюшев, В.В. & Короткий, А.А. (2020) Рекомендательные системы безопасности для риск-ориентированного подхода. *Безопасность труда в промышленности*. 5, 48–55.
- Петров, А.А. & Халафян, Р.М. (2013) Роль Международных И Внутригосударственных Рекомендательных Актов В Правовой Системе России. 305.
- Суходолов, А.П., Бычкова, А.М. & Ованесян, С.С. (2019) Журналистика с искусственным интеллектом. *Вопросы теории и практики журналистики*. 8 (4), 647–667.
- Харитонов, Ю.С. (2021) Правовое регулирование применения технологии искусственного интеллекта в военном деле: опыт России и Китая. *Журнал прикладных исследований*. 2 (1), 72–80.
- Харитонов, Ю.С. (2020) Правовые проблемы применения технологии искусственного интеллекта при принятии решений в военном деле. *Журнал прикладных исследований*. 2 (4), 54–57.
- Харитонов, Ю.С. & Ян, Т. (2022) Рекомендательные системы: выработка правовых подходов в России и Китае. *Предпринимательское право*. (2), 50–58.
- Центр контент-аналізу (2023) Оцінка вживання української та російської мов в українському сегменті соціальних мереж в жовтні 2023 року та порівняння з даними 2020 та 2022 років. 2023. Google Docs. <https://drive.google.com/file/u/0/d/1LX8mgQWCjT4-zgAJJA3Etiux79Ut->

[Tll/view?usp=sharing&fbclid=IwY2xjawLM989leHRuA2FlbQIxMABicmlkETBiajByNGladHJiWGxia2Q2AR5R8njvvuqYwmd8R0hr95A7fjCV24RkSLlgyvFaRaSTrc7B-](https://t.me/view?usp=sharing&fbclid=IwY2xjawLM989leHRuA2FlbQIxMABicmlkETBiajByNGladHJiWGxia2Q2AR5R8njvvuqYwmd8R0hr95A7fjCV24RkSLlgyvFaRaSTrc7B-)

[I4_a6W6Gbylyw_aem_hW04QYYsWtGaxiZVNkGLBA&pli=1&usp=embedded_facebook](https://t.me/I4_a6W6Gbylyw_aem_hW04QYYsWtGaxiZVNkGLBA&pli=1&usp=embedded_facebook) [Accessed: 28 June 2025].

- Шипунова, О.Д., Поздеева, Е.Г. & Евсеева, Л.И. (2021) Цифровые приложения и модели личности в контексте киберантропологии. *Социология*. (5), 234–239.
- ЭНЕРГО-И, Р.В.Т.И. (2022) Использование нейросетевых алгоритмов в энергетике для прогнозного регулирования. In: *Энерго-И Ресурсосбережение В Теплоэнергетике И Социальной Сфере: Материалы Международной Научно-Технической Конференции Студентов, Аспирантов, Ученых Учредители: Южно-Уральский государственный университет (национальный исследовательский университет)*. 2022 pp. 71–73. <https://elibrary.ru/item.asp?id=48310227>.