

# Biases, Epistemic Filters, and Explainable Artificial Intelligence

Sebastiano Moruzzi<sup>1,\*</sup>, Filippo Ferrari<sup>1</sup> and Filippo Riscica<sup>1</sup>

<sup>1</sup>Dipartimento delle Arti, Università di Bologna, via Azzo Gardino 23, 40126 Bologna, Italy

## Abstract

This paper examines the role of biases and epistemic filters in Explainable AI (XAI) and Generative AI (GenAI). The increasing integration of AI into social frameworks raises questions about human-machine interaction and governance within democratic systems. The study emphasizes the importance of incorporating social epistemology to address the complexities of AI-related epistemic questions, traditionally analyzed from an individualist perspective but now requiring a social approach. It highlights the need for transparent AI explanations to assist AI-based decision-making, considering stakeholders' biases and the effectiveness of XAI methods. The concept of epistemic filters—omission and discredit filters—is introduced to analyze how these biases impact AI outputs and user interactions. The paper also discusses the challenges of evaluating GenAI outputs and the necessity of prompt engineering skills, proposing a research agenda for employing epistemic filters to enhance XAI techniques. Ultimately, the goal is to ensure AI systems are not only technically accurate but also contextually appropriate, transparent, and fair, addressing both technical and cognitive biases.

## Keywords

bias, XAI, GenAI, epistemic filter

## 1. Introduction

The rise of digital technologies and the pervasive use of social media have profoundly altered our perception of society and communication. The radicalization of polarization dynamics in liberal democracies has been exacerbated by the impact of these technologies. The dissemination of opinions among communities that would otherwise have no way of communicating and aggregating, even if only virtually, has led to unprecedented modes of interaction within the political and social spheres.

The massive and rapid advent of artificial intelligence (AI) will inevitably impact these dynamics. Key questions arise: i) how will human-machine interaction integrate into the complex framework of social relations? ii) Can democratic systems effectively govern the formidable challenges AI presents to liberal democracies?

---

*HHAI-WS 2024: Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 10–14, 2024, Malmö, Sweden*

\*Corresponding author.

✉ [sebastiano.moruzzi@unibo.it](mailto:sebastiano.moruzzi@unibo.it) (S. Moruzzi); [filippo.ferrari16@unibo.it](mailto:filippo.ferrari16@unibo.it) (F. Ferrari); [filippo.riscica2@unibo.it](mailto:filippo.riscica2@unibo.it) (F. Riscica)

🌐 <https://www.unibo.it/sitoweb/sebastiano.moruzzi> (S. Moruzzi); <https://www.unibo.it/sitoweb/filippo.ferrari16> (F. Ferrari); <https://www.unibo.it/sitoweb/filippo.riscica2> (F. Riscica)

🆔 0000-0001-9189-2400 (S. Moruzzi); 0000-0001-6770-8124 (F. Ferrari); 0000-0002-0177-285X (F. Riscica)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

It is urgent to integrate the tools of philosophy, including social epistemology, into the study of AI. It is now a well-established trend in epistemology to address questions that concern not only the individual but also groups of individuals and complex social contexts. Central themes in epistemology, such as normative and metaphysical questions concerning knowledge, justification, and belief, which were typically addressed from an individualist perspective, are now increasingly analyzed from a social and non-ideal perspective, exploring how social interactions and systems contribute to epistemic outcomes [1, 2, 3, 4].

Social networks will be increasingly affected by human-machine interaction. Consequently, it will become more important to support AI-based deliberations with transparent and meaningful explanations of AI outputs. Explainable Artificial Intelligence (XAI) aims to provide insights into the predictions generated by machine learning models. These predictions have a variety of applications, each necessitating a distinct type of explanation. Much of the earlier XAI research focused primarily on creating new methods of explainability, rather than assessing whether these approaches effectively meet the needs and expectations of stakeholders [5]. It is crucial to design XAI methods that cater to stakeholders' needs and expectations. Furthermore, it is important to analyze how stakeholders' biases affect XAI-assisted decision-making [6]. An underexplored issue in relation to XAI-related biases is understanding how stakeholders' prejudices about the sources of evidence impact the effectiveness of the explanations provided by XAI methods.

The plan of this paper is as follows. Section 2 introduces the concept of epistemic filters, explaining their role as gatekeepers of evidence and distinguishing between omission and discredit filters. Section 3 discusses the growing need for explainability in Generative AI (GenAI) and how epistemic filters can help address biases in AI outputs and user interactions. Section 4 proposes a research agenda for GenAI and epistemic filters, emphasizing the need for interdisciplinary collaboration to operationalize these concepts effectively.

## 2. What are Epistemic Filters?

Whenever we interact with an AI system to get answers to our questions, we are engaged in inquiry. The term "inquiry" here refers to the complex practice of gathering, weighing, and assessing evidence aimed at forming, managing, and revising beliefs to acquire and share true information. AI systems have become increasingly integrated into our practices of inquiry, and to the extent that they act autonomously, they will also conduct their own inquiries (although the nature and dynamics of AI inquiries might be very different from human inquiry, such as their valuation of truth).<sup>1</sup>

---

<sup>1</sup>It is an interesting question whether the process leading an AI to answer a question can be properly identified as an inquiry. Perhaps it is not exactly an inquiry, but as far as we can identify some patterns similar to inquiry, such as giving reasons and adhering to evidence, we might call it a quasi-inquiry, or an "inquiry\*". The extent to which AI's inquiry\* differs from human inquiry deserves further scrutiny. Some features, such as being truth-oriented, might be missing. Some researchers have recently characterized generative AI systems such as ChatGPT using Frankfurt's concept of bullshitting, meaning that these AI systems act to persuade without regard for truth [7]. The issue of properly understanding meaning and truth for AI systems like ChatGPT might impact our application of epistemic filters to AI. If some norms constitutive of inquiry are absent in inquiry\*, it might still resemble inquiry as long as some functions of inquiry are preserved.

We can conceptualize the practice of inquiry as regulated by norms that determine when it is rational to form or revise beliefs given the available evidence. The normative structure of inquiry is constrained by two evidential norms. The first norm allows a subject to form a belief that  $p$  if and only if they have undefeated evidence for the belief that  $p$ . The second norm requires a subject to revise their belief that  $p$  if and only if they have an undefeated defeater for the belief that  $p$ . Epistemic filters function as gatekeepers of evidence. Given a set of total evidence, epistemic filters select which pieces of evidence will be relevant to assess whether belief formation or revision is appropriate. Thus, if a piece of evidence passes through the epistemic filters, it will fall within the range of application of the two evidential norms. Otherwise, it will not.

There are at least two types of epistemic filters: omission filters and discredit filters. Filtering by omission occurs when an evidential source is not accessible. The concept of reachability is used broadly, encompassing situations where agents do not have access to an evidential source (e.g., because the source is behind a paywall and the agent cannot afford it) and situations where agents may not understand the evidence. Filtering by discredit occurs when an evidential source is accessible but disregarded, perhaps because it is considered untrustworthy or intentionally misleading.

To clarify this analysis, consider the following example:

**Example 1.** *In a community, some people are virologists and epidemiologists who perform experiments and communicate their results on viruses, their diffusion patterns, and related vaccines. These scientists also intend to communicate their results to the public, including those without scientific training. Among the public, some distrust the scientific enterprise, particularly regarding vaccines (they believe this scientific research is inevitably unreliable due to economic interests). These individuals, who rely on alternative methods of inquiry and sources of information, can be referred to as group A. When members of group A hear new information about vaccines from a scientific source, such as the efficacy and safety results from clinical trials, they do not consider it reliable. Even if this information challenges their beliefs, they will not update their beliefs accordingly. For example, if new research demonstrates that a particular vaccine significantly reduces the incidence of a disease without serious side effects, anti-vaxxers will still reject this evidence.*

*However, there are also individuals who follow what the scientists say—group B—who initially share the same skepticism about vaccines as the anti-vaxxers. Yet, they are unaware of the new information from the scientists. If members of group B were to learn about the new scientific evidence, such as detailed results showing high vaccine efficacy and minimal adverse reactions, they would likely revise their beliefs. For instance, upon learning that a new vaccine has passed rigorous safety protocols and effectively protects against a virus, these individuals in group B might change their stance and support vaccination.*

The concept of epistemic filters helps make sense of this situation without necessarily imputing irrationality to members of groups A and B. Group A's selection of sources operates through a discredit filter. Even if they become aware of new information, their epistemic discredit filter deems it appropriate to ignore it. Conversely, for people in group B, it is appropriate to share the same belief as post-inquirers because they are not aware of the new information, making this a case of filtering by omission.

Recent research in social epistemology has argued that inquiry must always be understood as situated in a context characterized by multifaceted considerations, ranging from our conception of the natural world to the social values we embrace [8, 9, 10]. Insofar as epistemic filters encode social values, inquirers shape their space of inquiry accordingly by adopting epistemic filters for forming and assessing their reasons for acting, as well as for forming and revising their beliefs.<sup>2</sup>

If members of these groups can also be AI systems, we can conceptualize the dynamics of interaction between these groups as being modeled by epistemic filters that encompass not only human-human interaction but also human-machine interaction and machine-machine interaction.

However, interactions involving AI systems require further scrutiny, as it is often unclear what reasons underlie their outputs. Our hypothesis is that applying the concept of epistemic filters to AI systems can help improve the explainability of their outputs.

Having introduced the concept of epistemic filters, we will now argue that epistemic filters can be a useful conceptual tool for XAI.

### 3. XAI, GenAI, and Epistemic Filters

The need for explainability in Generative AI (GenAI) is growing as humans control and customize its outputs. GenAI blurs the line between users and developers, allowing non-programmers to create applications using models like OpenAI's GPT-4. Prompt engineering has become a crucial skill, and explainable AI (XAI) is essential to support it. Users must learn to manage outputs, handle limitations, and mitigate risks. Therefore, stakeholders must understand GenAI to develop solutions that meet their needs.

Schneider [11] has identified several challenges that XAI faces with GenAI. Users and researchers cannot access the internals and training data of commercial GenAI models, limiting XAI approaches. Explanations should focus on the model's impact on humans during interactions. Understanding AI is harder as models grow, using larger data and external tools. GenAI outputs involve many decisions, leading to investigations of properties like tone and style. Evaluating explanations is challenging without benchmarks for XAI methods. GenAI models are used by diverse users, unlike pre-GenAI systems. They can produce offensive or biased content, and explanations might be poorly phrased. GenAI also suffers from hallucinations and limited reasoning, affecting self-explanations.

To tackle these challenges, new desiderata have emerged for XAI. One prominent aspect is *lineage*, which ensures that model decisions can be tracked to their origins:

[lineage] refers to tracking and documenting data, algorithms, and processes throughout the lifecycle of an AI model. It is highly relevant for accountability, transparency, reproducibility, and, in turn, governance of artificial intelligence [142]. It concerns the “who” and the “what”, for example, “Who provided the data or made the model?” or “What data or aspects thereof caused a decision?”. While the

---

<sup>2</sup>Following Nguyen [8], we can characterize an epistemic bubble as a social epistemic structure in which agents have their inquiry involving omission filters, while an echo chamber is a social epistemic structure in which agents have their inquiry involving discrediting filters. We leave aside discussion of these social structures, though we think they are further elements to consider in the context of XAI.

latter is a well-known aspect of XAI, as witnessed by sample-based XAI techniques, the former has not been emphasized significantly in the context of XAI. [37] set forth data traceability as a requirement in the context of Machine Learning Operations (MLOPs) for XAI in industrial applications. The need for lineage emerges as GenAI supply chains get more complex often involving multiple companies [146] rather than just a single one. Furthermore, multiple lawsuits have been undertaken in the context of generative AI, for example, related to copyright issues [50]. Regulators have also set stringent demands on AI providers [36]. Thus, employing GenAI poses legal risks to organizations. In turn, ensuring lineage-supported accountability can serve as a risk mitigation. [11, §3.3]

Systematic biases can be hidden in the lineage of a GenAI model and impact interactions with users. At the same time, users' biases can also impact these interactions, and unless a bias check is done, two different types of biases can undermine the efficiency of the interaction between GenAI and users.

To illustrate: suppose part of the dataset related to vaccine safety is grounded on a dataset that includes blog posts of vaccine skeptics who distrust the scientific enterprise on viruses and vaccines and rely on alternative methods of inquiry and sources of information. A user who trusts scientific evidence on clinical trials for vaccines might inefficiently interact with GenAI if they are not aware that the lineage of some outputs is grounded in data from skeptics.

These kinds of systematic biases can be aptly explained by the presence of epistemic filters.

We envisage at least two ways in which the concept of epistemic filters can be operationalized to tackle two dimensions of explanation properties [11, §4.1.1]:

- **Explanations of single input-output relations:** Making explicit the relevant epistemic filter adopted by the GenAI system can help explain how the input produced a certain output.
- **Explaining the entire interaction:** Understanding how relevant epistemic filters affect the dynamics of communication between an AI and a human who employs the AI to address and solve a certain task.

The epistemic filter model can help analyze and address the bias problem in GenXAI by addressing issues related to these two explanation properties. This model enables us to examine how user and AI biases can influence the interpretation of input-output relations and the dynamics of interaction.

With respect to interactions, which are a key element of the user experience with GenAI, Schneider [11] notes:

Interaction dynamics are influenced by a series of technical (such as model behaviour, including classical performance measures but also latency, user interface, etc.) and non-technical factors (such as human attitudes and policies). As such, human-AI interaction cannot easily be associated with one scientific field but is inherently interdisciplinary. Explainability, which aims at understanding AI technology, should focus on how technical factors related to model behaviour impact

the interaction. While many existing works touch on the subject, the change in interactivity brought along by prompting due to GenAI is not well understood.

In the following, we provide two contexts in which epistemic filters can help implement GenXAI techniques. To do that, we refer to some of the XAI techniques listed in the taxonomy of XAI techniques provided by Schneider [11].

For each context, we imagine the role that epistemic filters might play in relation to existing XAI techniques. We call these hypotheses “proto-operationalization” since they are just a sketch of how epistemic filters could be operationalized for GenXAI.

### 3.1. Identifying Epistemic Filters in XAI Interpretations

Every user brings a set of epistemic filters when interacting with AI systems. These filters can significantly affect how AI explanations are perceived and understood. For example, if a user inherently distrusts machine learning models due to previous negative experiences or a lack of familiarity with the technology, they might interpret explanations with skepticism or a predisposed bias against the conclusions drawn by the AI. Epistemic filters can help identify these biases and understand how they might distort the user’s interpretation of the AI’s explanations.

**Proto-operationalization for AI:** By examining the training data relevant to the topic of interaction, we can measure the impact of a specific training sample on the output based on the epistemic filter grounded in the sample.

**Example:** If the sample includes blog posts from anti-vaxxers, the background assumption of the unreliability of scientific evidence on clinical trials for vaccines becomes a key feature to explain the output.

**Proto-operationalization for users:** By eliciting the background assumptions that express the epistemic filters of the user relevant to the topic of interaction between user and AI, we can explain the dynamics of communication between human and AI.

**Example:** If the AI output includes a type of evidence that the user distrusts because they adopt a certain epistemic filter (e.g., alternative medicine strategies), then the epistemic filter will explain why such an interaction will have a certain dynamic (the user will ignore the prompt and seek alternative explanations).

### 3.2. Tailoring Explanations to User-Specific Epistemic Filters

By understanding the specific epistemic filters through which different users view AI explanations, developers can tailor explanations to be more effective and comprehensible to various user groups. This involves adjusting the complexity, format, and content of explanations based on the users’ epistemic backgrounds. For instance, a highly technical explanation involving detailed model parameters and algorithms might be appropriate for a data scientist but could be entirely opaque to a layperson, who may require a more qualitative, simplified explanation. This point is particularly relevant when AI is used in the judicial system and in the exercise of public authority [12, 13].

**Proto-operationalization for AI and users:** By employing probing-based methods such as concept-based explanations [11], we can uncover the concepts relevant to the input information

topic. The presence of certain concepts indicates the conceptual level employed by the AI in relation to the topic of interaction and provides a measure of the conceptual distance between user and AI. Such distance might explain the presence of omission filters that impact the interaction.

**Example:** If the concepts in the AI input relate to sophisticated technical knowledge of biomolecular medicine, but the user's prompts reflect only non-technical concepts of medicine, the distance between these concepts is evidence of an omission filter that explains why AI outputs are not effective for the user in the interaction.

To sum up, by making users aware of their own epistemic filters and how these may influence their understanding, AI developers can encourage a more critical and informed engagement with AI systems. This involves not only explaining what the AI system does but also educating users about common cognitive biases and demonstrating how these might influence their interaction.

Ultimately, the application of epistemic filters can enhance the overall effectiveness of explainable AI. By ensuring that explanations are not only technically accurate but also contextually appropriate and understandable for different users, AI systems can become more transparent and fair. This approach addresses the bias problem by acknowledging that bias in AI is not just a technical issue but also a perceptual and cognitive one.

#### **4. Research Agenda for GenAI and Epistemic Filters**

To effectively employ epistemic filters in GenXAI, we need to work on the operationalization of the concept of epistemic filters. This requires more interdisciplinary collaboration between philosophers and computer scientists.

We envision at least two research directions that can be pursued:

1. We need methods to infer the presence of epistemic filters from a dataset. The research hypothesis is that by analyzing the provenance of the dataset and the information about the identities of the data producers, we can reliably infer which epistemic filters are present.
2. Epistemic filters have been formalized using the bounded-confidence model of opinion dynamics (the Hegselmann–Krause model with truth parameters, see [14]). The research hypothesis is that such a model of opinion dynamics can be used to explain the interaction between users and AI in the context of GenXAI.

#### **5. Conclusions**

The epistemic filter model offers a sophisticated tool for understanding and improving how explanations in AI are generated and received. It recognizes that overcoming bias in AI is not merely about adjusting data or algorithms but also about addressing the human factors that influence how AI systems are perceived and used. This dual focus is crucial for the development of truly effective and trustworthy AI systems.

## Acknowledgments

Thanks to Chiara Natali for organizing the Frictional AI workshop and for providing a wonderful opportunity to connect with an interdisciplinary community of researchers on AI.

## References

- [1] E. Begby, *Prejudice: A Study in Non-Ideal Epistemology*, Oxford University Press, Oxford, 2021.
- [2] S. Fuller, *Social Epistemology*, Indiana University Press, 2002.
- [3] A. Goldman, C. O'Connor, *Social epistemology*, *The Stanford Encyclopedia of Philosophy* (2021). Winter 2021 Edition, Available at: <https://plato.stanford.edu/archives/win2021/entries/epistemology-social>.
- [4] R. B. McKenna, *Non-Ideal Epistemology*, Oxford University Press, Oxford, 2023.
- [5] M. Langer, D. Oster, T. Speith, L. Kästner, K. Baum, H. Hermanns, E. Schmidt, A. Sesing, What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research, *Artificial Intelligence* 296 (2021) 103473.
- [6] A. Bertrand, R. Belloum, J. Eagan, W. Maxwell, How cognitive biases affect xai-assisted decision-making: A systematic review, in: *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, Oxford, United Kingdom, 2022, pp. 78–91.
- [7] M. T. Hicks, J. Humphries, J. Slater, Chatgpt is bullshit, *Ethics and Information Technology* 26 (2024) 1–10. doi:10.1007/s10676-024-09775-5.
- [8] C. T. Nguyen, Echo chambers and epistemic bubbles, *Episteme* 17 (2020) 141–161. doi:10.1017/epi.2018.32.
- [9] F. Ferrari, S. Moruzzi, Enquiry and normative deviance, in: S. Bernecker, A. K. Flowerree, T. Grundmann (Eds.), *The Epistemology of Fake News*, Oxford University Press, New York, NY, 2021, pp. 109–133.
- [10] F. Ferrari, S. Moruzzi, Post-enquiry and disagreement. a socio-epistemological model of the normative significance of disagreement between scientists and denialists, *Social Epistemology* 37 (2023) 177–196.
- [11] J. Schneider, Explainable generative ai (genxai): A survey, conceptualization, and research agenda, *ArXiv abs/2404.09554* (2024). URL: <https://api.semanticscholar.org/CorpusID:269149593>.
- [12] L. Beckman, J. H. Rosenberg, K. Jebari, Artificial intelligence and democratic legitimacy. the problem of publicity in public authority, *AI and Society* (Forthcoming) 1–10.
- [13] A. Deeks, The judicial demand for explainable ai, *Columbia Law Review* 119 (2019) 1829–1850.
- [14] F. Ferrari, S. Moruzzi, F. Riscica, Modelling epistemic filters., in: S. M. Popa (Ed.), *Vices of the Mind: Mis/Disinformation and Other Epistemic Pathologies*, Cambridge University Press, forthcoming.