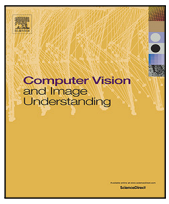
Computer Vision and Image Understanding 248 (2024) 104142

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Embedding AI ethics into the design and use of computer vision technology for consumer's behaviour understanding

Simona Tiribelli [a], Benedetta Giovanola [a], Rocco Pietrini [b,*], Emanuele Frontoni [a], Marina Paolanti [a]

[a] Department of Political Sciences, Communication and International Relations, University of Macerata, Via Don Minzoni 22/A, 62100 Macerata, Italy
[b] Università Politecnica delle Marche, VRAI - Vision Robotics and Artificial Intelligence Lab, Dipartimento di Ingegneria dell'Informazione (DII), 60131 Ancona, Italy

ARTICLE INFO

Keywords:
Human behaviour analysis
Artificial intelligence
AI ethics
Retail environment

ABSTRACT

Artificial Intelligence (AI) techniques are becoming more and more sophisticated showing the potential to deeply understand and predict consumer behaviour in a way to boost the retail sector; however, retail-sensitive considerations underpinning their deployment have been poorly explored to date. This paper explores the application of AI technologies in the retail sector, focusing on their potential to enhance decision-making processes by preventing major ethical risks inherent to them, such as the propagation of bias and systems' lack of explainability. Drawing on recent literature on AI ethics, this study proposes a methodological path for the design and the development of trustworthy, unbiased, and more explainable AI systems in the retail sector. Such framework grounds on European (EU) AI ethics principles and addresses the specific nuances of retail applications. To do this, we first examine the VRAI framework, a deep learning model used to analyse shopper interactions, people counting and re-identification, to highlight the critical need for transparency and fairness in AI operations. Second, the paper proposes actionable strategies for integrating high-level ethical guidelines into practical settings, and particularly, to mitigate biases leading to unfair outcomes in AI systems and improve their explainability. By doing so, the paper aims to show the key added value of embedding AI ethics requirements into AI practices and computer vision technology to truly promote technically and ethically robust AI in the retail domain.

1. Introduction

In the last decades, a growing corpus of literature and guidelines has been developed for the ethical, human-centred, and trustworthy use of Artificial Intelligence (AI) systems and machine learning (ML) and deep learning (DL) algorithm-based technology (Jobin et al., 2019; Corrêa et al., 2023). In particular, such efforts spurred out from a series of risks and detrimental phenomena unveiling how AI systems when designed and deployed without embedding ethical and societal considerations can both intentionally and/or inadvertently harm specific individuals, groups, and societies. Risks and concerns highlighted span over people's privacy infringements and personal data misuses, phenomena of tech surveillance for human behaviour's manipulation, and unfair and biased AI-based outcomes used in support of human decision-making — just to mention a few (Giovanola and Tiribelli, 2022). Many of such risks are particularly linked to the disruptive capacities of such systems to process huge amounts of data and discover precious patterns and correlations on how things are "likely to be" in the future and how is likely "we will behave", that is, their predictive huge potential linked to their probabilistic nature (Tiribelli, 2024). Such potential rapidly found a fertile ground in the retail sector, where the capacity to deeply understand and even predict human agency and particularly consumer behaviour and decision-making can draw the fine line between vendors and products that succeed and those that instead fail (Fildes et al., 2022; Kliestik et al., 2022). Unsurprisingly, the retail domain has assisted to the rapid implementation of a number of novel AI techniques with a high rate of success in terms of retail management's efficiency and retail strategies' productivity (Pascucci et al., 2022).

However, while such techniques are becoming increasingly pervasive in the domain of "onlife" (Floridi, 2014) or "phygital" retail, the ethical considerations underpinning their trustworthy deployment have been poorly explored to date. This lack is problematic. Indeed, many of the ethical and societal risks and challenges pointed out in the AI ethics scholarship turn out to be particularly pressing in the AI-empowered

* Corresponding author.
E-mail address: r.pietrini@staff.univpm.it (R. Pietrini).

https://doi.org/10.1016/j.cviu.2024.104142
Received 13 May 2024; Received in revised form 22 July 2024; Accepted 2 September 2024
Available online 4 September 2024

retail experience and specifically for the individuals at the centre of the retail environment — who may end up to be vulnerable to such risks if unconsidered. Furthermore, in the long term, over-looking such risks and considerations might lead people refusing the use of AI in the retail space, producing what is known as the "AI opportunity loss" effect, that is, the underuse of AI, caused by its intentional or accidental misuse. Therefore making AI techniques for retail trustworthy and human-centred by embedding ethical considerations and requirements by design turns out to be of critical importance. However, this issue poses a particular research question: what do AI ethical principles and requirements mean and entail for the trustworthy design and use of AI for human behaviour analysis in retail ecosystems?

Indeed, in retail environment, understanding consumer behaviour is critical (Rossi et al., 2021; Ferracuti et al., 2019). Recent innovations in RGB-D (depth sensing) camera technology have improved our ability to monitor and analyse how shoppers navigate and interact within stores, despite challenges such as occlusions, dynamic backgrounds and varying lighting conditions (Paolanti et al., 2020). These cameras, used in a top-view configuration, have been particularly effective in maintaining privacy and reducing data complexity by focusing on movement patterns rather than individual identities (Martini et al., 2020; Frontoni et al., 2019). AI-based systems have been employed to count the number of people passing or stopping in the camera area, perform top-view re-identification and measure shopper–shelf interactions from a single RGB-D video flow with near real-time performances (Paolanti et al., 2020). However, while the ethical and societal challenges posed by such systems are sector-specific, to date the ethics and policy guidelines to tackle them tend to be high-level and henceforth ineffective in providing actionable guidelines to engineers and stakeholders working in the domain of retail. To address these gaps, the goal of this paper is to (i) provide an up-to-date detailed analysis of the most advanced AI/CV techniques for consumer behaviour understanding in the retail sector and (ii) highlight the key ethics issues to be considered in designing and using them to ensure their development and deployment in compliance with AI ethics principles and requirements for the trustworthy AI at the European (EU) level. In particular, we aim to address the following research questions: can AI ethical principles support the trustworthy design and use of consumer applications in retailing? If so, how, and what do they mean in the field of AI for human behaviour analysis in retail?

To do so, considering the high-levelness of many ethical frameworks of principles claimed in the scholarship in AI ethics, we considered agreed-upon AI ethics requirements (see, in particular, HLEG-AI 2019 (AI, 2019) and ALTAI 2020[1]) in a context-sensitive way, that is, by matching high-level considerations with the peculiarities emerging from the context analysed, with a specific focus on the ethical requirements of fairness and explainability. Put it differently: we merge a top-down approach grounded on up-to-date AI ethics literature and policy documentation with technicalities and specificities emerging from the technical literature in the field of AI and CV for the retail sector. We present a tailored AI ethics framework that emphasizes the integration of bias detection and strengthened explainability, with the aim of promoting concrete and perceived trustworthiness in retail AI systems. Through a comprehensive analysis and evaluation of the VRAI framework (Paolanti et al., 2020) using advanced explainability metrics, we identify and address significant gaps in its current implementation. The VRAI framework has been instrumental in advancing retail analytics applications using three AI models to perform simultaneous people counting, top view re-identification, and shopper-shelf interaction analyses. Our research highlights the need to embed ethical considerations and bias mitigation strategies into AI development, ultimately leading to the creation of fairer and more understandable AI solutions for the retail sector. Put it differently: we merge a top-down approach grounded on up-to-date AI ethics literature and policy documentation with technicalities and specificities emerging from the technical literature in the field of AI and CV for the retail sector. By doing so, we fill a gap in the scientific literature on AI and ethics by providing an ethical compass to navigate ethical considerations and hence guide diverse stakeholders in the design and use of trustworthy AI techniques for human-centric retail. Furthermore, our research shows how to embed ethical considerations and bias mitigation strategies into AI development, ultimately leading to the creation of fairer and more understandable AI solutions for the retail sector. From this perspective, our research paves the way to facilitating diverse stakeholders using AI in retail in complying with the numerous and often opaque AI ethics standards and guidelines increasingly binding at the international level.

The main contributions of the paper can be summarized as follows: (i) a methodology for integrating high-level ethical frameworks and benchmark principles from AI ethics scholarship with the specific needs and contexts of the retail sector. This approach aims to make ethical guidelines more actionable and relevant for engineers and stakeholders in the retail domain. (ii) Context-sensitive ethical framework. This framework aims to guide stakeholders in developing AI systems that are not only technologically effective but also ethically robust. (iii) The application of the framework to extend the VRAI framework towards AI ethics principles. (iv) An ethical guide for the various stakeholders involved in the design and use of AI in retail. It provides a roadmap for navigating ethical considerations and ensuring the development of human-centred, trustworthy computer vision solutions.

The paper is structured as follows: Section 2 provides a thorough review of the existing literature on AI for human behaviour analysis especially in the retail domain and identifies the key gaps that our research aims to fill with reference to the application of ethical frameworks. Section 3 outlines the Ethical Principles for Trustworthy CV in understanding consumer behaviour. Section 4 summarizes the EU vision for trustworthy AI and particularly the prominent ethical principles and requirements for the development of technically and socially robust AI systems highlighted in the EU context. Section 5 proposes strategic measures to addresses the challenges around explainability and the presence of bias identified in the VRAI framework. In Section 6 enhancements aimed at strengthening the trustworthiness of the system are proposed. The final section (Section 7) summarizes the key findings and contributions of the paper. It also discusses the implications for future research and suggests how this work can be extended and refined to further support the ethical use of AI in retail and other sectors.

2. Related works

This section provides a comprehensive review of the existing literature on the ethical considerations about the analysis of human behaviour, by exploring works and foundational studies that have framed the general discourse on AI ethics, particularly those that address privacy, consent, and data security concerns.

In Suarez et al. (2023), the authors undertook a thorough review and analysis of ethical decision-making models found in peer-reviewed publications in behaviour analysis and various allied health fields. Their review uncovered 55 different ethical decision-making models presented in 60 scientific articles spanning seven primary professions, such as medicine and psychology, and 22 sub-fields, including dentistry and family medicine. Consensus-based analysis revealed nine behaviours commonly recommended by these models, with the majority (n = 52) sequentially organizing these behaviours and less than half (n = 23) incorporating a problem-solving element. All nine identified steps closely align with those outlined in the Code of Ethics for Behaviour Analysts, to be published by the Behaviour Analyst Certification Board in 2020, suggesting broad professional agreement on the behaviours likely to be integral to ethical decision-making.

[1] https://altai.insight-centre.org.

Contreras et al. (2021) argued that evidence-based practice (EBP) in applied behaviour analysis (ABA), as defined by Slocum et al. (2014), provides a structured approach to enhancing ethical decision-making processes. In this paper, they emphasized the importance of ethical decision-making in ABA practice and introduced and reviewed the EBP approach in ABA. The relationship of EBP in ABA to the ethical standards of the Behaviour Analyst Certification Board is highlighted, along with suggested actions for behaviour analysts to continually enhance ethical decision-making.

Similarly, in the context of human behaviour analysis, in Wilkenfeld and McCarthy (2020) the authors sought to determine what makes the treatment of Autism Spectrum Disorder (ASD) both effective and ethical. They claimed that a widely used method of Applied Behaviour Analysis (ABA), often regarded as the superior approach to treating ASD, systematically violates basic bioethical principles. Furthermore, the purported benefits of this treatment not only fail to address these violations but tend to exacerbate them. Although concerns about ABA have been raised by autism advocates for some time, these warnings have largely gone unheeded and ABA remains a common recommendation, often aggressively promoted to parents of autistic children. In particular, it has been argued that the use of ABA violates the bioethical principles of justice and nonmaleficence and critically compromises the autonomy of both children and, in cases of forceful advocacy, parents.

Kelly et al. (2021) examined the ethical principles of their organization based on the certified behaviour analysts who are required to adhere to the ethical rules established by the Behaviour Analyst Certification Board® (BACB®), known as the Professional and Ethical Compliance Code for Behaviour Analysts. They explored how behaviour analysts can use them to make both clinical and ethical decisions, and how to overcome the challenges of dissemination. Ethical guidelines that are not based on clear principles can present barriers to dissemination for behaviour analysts who need to communicate the ethical standards of their field to colleagues and stakeholders from non-behavioural backgrounds. This article describes how their organization, the BACB, has developed a set of guiding ethical principles to complement the BACB Code. These principles assist members in making ethical decisions and help them to communicate their organizational values effectively.

Moreover, the principles of human behaviour analysis extend beyond clinical applications and have been effectively utilized in various other fields, including retail and organizational behaviour.

In the retail domain, there is a significant gap in the literature on ethical guidelines for the analysis of human behaviour, particularly in the context of computer vision technologies. A pioneering effort to address this gap was made by Anica-Popa et al. in their study, which aimed to explore the practical benefits and risks associated with AI applications in retail (Anica-Popa et al., 2021). They aimed to use these findings to develop a conceptual framework for integrating AI technologies into retail information systems. To achieve this, Popa et al. conducted a systematic review of recent literature focusing on AI implementations in the retail industry. The findings from this review helped to establish a conceptual framework. Their research uncovered several sophisticated AI solutions that offer numerous benefits but also pose certain risks, in different segments of the retail value chain. This chain, abbreviated as CECoR, includes improving the customer experience (CE) through technologies such as virtual agents, reducing costs (Co) through innovations such as smart shelves, and increasing revenues (R) through targeted product recommendations and personalized promotions. The conceptual framework is centred on customer profiles and provides detailed recommendations for implementing AI in retail environments, guided by the CECoR principles. Their findings are intended to be useful for both practitioners and researchers in the field, providing practical examples of the benefits, challenges and risks of AI technologies. The CECoR framework is intended to serve as a valuable tool for retailers and AI professionals alike, providing clear guidelines for initiating and managing AI integration projects within an organization's information systems.

Given that the ethical implications of using AI tools have not been adequately addressed in the current state of the art, in the next sections we consider recent advances in computer vision applications to gain a deep understanding of consumer behaviour and zoom in on two major ethical requirements for trustworthy AI. In the following section, we first list the core AI ethics principles and requirements for the trustworthy development, deployment and use of AI technologies as stressed and largely shared in the EU context.

3. Ethical principles for trustworthy CV for consumer's behaviour understanding

Embedding AI ethics principles into the development and deployment of specific AI systems requires (i) *explaining* what benchmark AI ethics principles entail and (ii) *unpack* them into categories and criteria to enable their enforcement and their comprehension by engineers and all the stakeholders involved, called to assess/approve such systems, or subject to specific AI techniques (both retailers and consumer). To do so, we rely on the "Ethics Guidelines for Trustworthy AI", developed by the High-Level Expert Group on AI set up in 2019 by the EU Commission (EC) and on the conceptual tool (*Assessment List for Trustworthy AI*) proposed for supporting AI developers and deployers in their concrete operationalization in specific sectors (see ALTAI 2020). The ethical requirements proposed by EC for Trustworthy AI aim to ensure the respect of 5 main ethical principles shared extensively in the scholarship in AI ethics (Jobin et al., 2019):

1. Benevolence;
2. Non-Maleficence;
3. Autonomy;
4. Justice & Fairness;
5. Explicability.

These ethical principles prescribe what should be done to use AI to benefit people and society, namely, in a trustworthy manner: using AI for good (1) and to minimize harm (2); respecting human autonomy and freedom of choice (3); ensure that AI does not discriminate in access and benefits due to unfair biases (4); and ensure AI systems are explainable and intelligible to those that are subject to them (5). In this work, we particularly focus on the two most prominent AI ethics principles and themes: (4) Justice & Fairness and (5) Explicability. Drawing on benchmark scientific scholarship in AI ethics, in the next section, we propose (a) a qualitative framework for unpacking and advancing fairness in AI systems for behaviour analysis through bias detection and prevention/mitigation; (b) a qualitative framework for assessing explainability by dimension and level; and (c) a few AI ethics metrics to assess and quantify explainability from a stakeholder perspective to boost the development and implementation of trustworthy AI systems in the retail domain.

4. Methodology

In this section, we present a detailed approach to embedding AI ethics in the design and use of computer vision technology for to understand consumer behaviour. The methodology is based on two conceptual frameworks we elaborate drawing on benchmark scientific scholarship in AI and ethics we propose to address critical concerns in AI ethics: Bias Detection and Explainability. Here are the detailed steps and processes involved:

- *AI ethics framework for bias detection:* This component focuses on identifying potential biases in the data input, model training and output phases. It outlines a procedure for systematically examining datasets for representativeness and inherent bias, using statistical and machine learning techniques to detect anomalies.

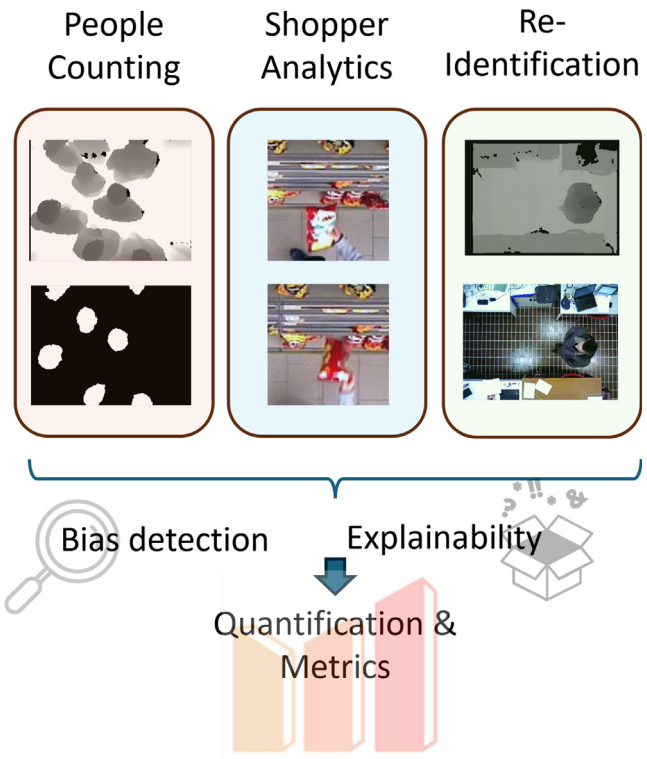


Fig. 1. Workflow of human behaviour analysis and AI ethics for retail.

- *AI Ethics Framework for Explainability:* This component ensures explicability of AI decisions, which is crucial for systems that interact directly with consumers. It incorporates methods to understand and enhance model interpretability and communicate AI decision processes clearly to stakeholders.
- *Quantification and Metrics:* Quantified indicators such as explanation fidelity, clarity scores, and user understanding metrics are integrated to measure how effectively the explainability interventions are understood by non-technical users.

Fig. 1 schematically depicts the ethical evaluation process for human behaviour understanding methodologies. This section is methodically divided into three integral components: "AI ethics framework for bias detection", "AI Ethics Framework for Explainability", and "Quantification and Metrics", each of which plays a pivotal role in our research.

4.1. AI ethics framework for bias detection

In addressing the ethical challenges associated with the use of AI in the retail sector, it is essential to critically examine certain components of the AI ethics framework, in particular those related to bias detection. These aspects are critical to ensuring that AI systems do not perpetuate discrimination or produce harmful effects. The study of bias in AI systems focuses on identifying and mitigating unintentional biases that may arise during the data processing or algorithmic decision-making phases. This is particularly relevant in the retail sector, where AI-driven recommendations or decisions may influence consumer choices and perceptions. We explore how the AI ethics principle of justice and fairness, as outlined by the EC High-Level Expert Group, can be operationalized starting with a framework for detect and correct a wide range of biases drawing on benchmark scholarship in the field (Giovanola and Tiribelli, 2023; Migliorelli et al., 2023; Mehrabi et al., 2021; Suresh and Guttag, 2021; Olteanu et al., 2019). This involves developing methods that not only assess the fairness of algorithmic outcomes but also ensure that all demographic groups are fairly represented and treated by AI systems. Tables 1 and 2 serve as a reference for data scientists, AI developers and policymakers involved in the design and deployment of AI systems. It acts as a conceptual compass to ensure that potential biases are identified and addressed in the early stages of AI system development. By addressing these biases, developers can improve the ethical and practical outcomes of AI applications, making them more equitable and trustworthy. By providing this structured breakdown, users can more effectively identify specific biases relevant to their projects and implement recommended actions to mitigate the risks associated with biased data. The table not only raises awareness but also empowers AI practitioners to proactively implement more ethical practices.

4.2. AI ethics framework for explainability

To ensure that AI systems are used responsibly in retail environments, it is essential to incorporate a robust explainability framework. This framework not only increases intelligibility of AI systems but also fosters trust among users by clarifying how AI decisions are made. The AI Ethics Framework for Explainability is proposed to address these needs by detailing the processes and methodologies that make AI actions understandable to both technical and non-technical stakeholders. The framework is structured around a multi-dimensional approach to explainability, grounded on Cabitza et al. (2023) and Ding et al. (2022), articulated through several dimensions, each of which addresses different aspects of how AI systems process data and make decisions. These dimensions are:

- Computational Explainability: Understanding the mechanical processes (the how) through which AI algorithms produce results.
- Justificatory Explainability: Clarifying what is the phenomenon causing the AI outputs.
- Informative explainability: Communicating what the outputs entail in practice.
- Cautionary explainability: Indicating the level of confidence and potential uncertainties associated with the AI's outputs.

In addition, the framework considers explainability at different levels of AI system operation, from global to local and semi-local, each providing a different depth of insight into the system's functionality (Cabitza et al., 2023). This layered approach ensures that explanations are available not only at an overall system level but also at more granular levels of individual decisions or model behaviours. These dimensions and levels of explainability are comprehensively outlined in Table 3, which serves as a guide for implementing the framework in practical AI applications in retail environments. Table 3 provides definitions and examples for each dimension and level, providing a clear roadmap for developers and stakeholders to improve the accountability of their AI systems.

By systematically applying this framework, AI developers can ensure that their systems are not only effective but also adhere to ethical standards that promote fairness, accountability and explainability, as essential requirements for maintaining public trust in AI technologies.

4.3. Quantification and metrics

Drawing on prominent scholarship on AI ethics for explainability (Hoffman et al., 2018), we have detected quantifiable qualitative metrics for understanding the explainability of these systems from a stakeholder perspective. These metrics are crucial for assessing how understandable and accessible the AI's decision-making processes are to users. These metrics are categorized into four main groups:

Table 1
Bias in data input for enacting fairness in AI systems. First part.

Type of bias	Description	Recommended Action (RA)
Population target bias	It arises when the characteristics of the effective users of the system differ from the intended target population (i.e., mismatch between effective user population of the AI system and prospect population targeted by design), leading to inaccurate outcomes.	Define carefully the target users and collect data as much heterogeneous and representative of different target users' characteristics as possible.
Missing data bias	Occurs due to datasets being quantitatively or qualitatively limited, affecting accuracy and generalizability of the systems' output.	Conduct thorough analyses to ensure full representativeness from qualitative and quantitative standpoint of the data collection sample used to train and test the model.
Minority bias	It occurs when data lacks sufficient representation of minority groups, affecting the model's accuracy and possibly producing discriminating outputs against them.	Ensure specific attention to minority groups and relative representation at the data entry and training level of the system, by using synthetic data if needed.
Informativeness bias	This is due to the availability of features that are less informative to render an accurate prediction for a group; for example, identifying face characteristics from an image of a user with dark skin may be more difficult.	Detect in advance and enhance by design less informative features to improve prediction accuracy especially for minority groups.
Temporal bias	Results from using outdated data that do not reflect current trends or behaviours leading to inaccurate outputs possibly leading to perpetuation of historical inequalities.	Continuously update the dataset to reflect current realities and trends with particular attention to check they do not capture behaviour and practices abandoned or rejected in the present societies.
Socio-behavioural context bias	It is due to variations in user behaviour across different platforms and socio-relational contexts, leading models to struggle to handle the complexity and diversity of real-world scenes, bearing poor robustness.	Consider dataset provenance and possible target users' behaviour variations across different platforms and socio-relational context.
Self-selection bias	Occurs when data collection is based on a self-selecting group (e.g., groups that decide to participate to the model's test), which may be limited in representativeness.	In the design phase include in the data sample participants both self-selected and randomly selected.
Historical bias	It refers to biases historically embedded in society and embedded in the data entry and training dataset.	Audit dataset to assess whether they contain correlations reflecting historical.bias and inequalities.

Table 2
Bias in data input for enacting fairness in AI systems. Second part.

Type of bias	Description	Recommended Action (RA)
Label bias	It occurs when labels assigned during data annotation are subjective, inconsistent or inaccurate, which may amplify data bias regarding gender, age, ethnicity, skin colour, etc.	Ensure labels are more objective as possible, clear, consistent, and interpreted across different groups, also according consensus voting sessions; this requires high heterogeneity (ensure the diversity of annotators) and competency in the design team.
Omitted variable bias	This bias arises when crucial variables are omitted from the model both in self-supervised or semi-supervised learning, leading to inaccurate predictions.	Include all relevant variables, consult stakeholders and experts during variable selection.
Aggregation bias	Arises when assumptions about individuals are based on aggregated data, potentially misleading.	Focus on data granularity and avoid assumptions only based on aggregated data alone.

Table 3
AI ethics framework for assessing and implementing explainability.

Multidimensional explainability		
Dimension	Definition	Example in retail
Computational	How the algorithm produces any output O.	Detecting shopper-shelf interaction by comparing pair of images of hands approaching/leaving the shelf.
Mechanistic	Why the algorithm produced the output O.	Because a product is detected leaving the shelf is more likely to have a positive interaction than not.
Justificatory	Why the output O is correct.	Because the product is missing from the shelf after the interaction.
Informative	What the output O means.	A sellout estimation can be done.
Cautionary	The degree of uncertainty behind the output O.	Accuracy metrics on image classification.
Explainability by level		
Level	Definition	Details/Example
Global	Provides a global understanding of the AI model's logic.	Especially hard to render for black box models; provides an understanding of outcome distributions.
Local	Segmenting the solution space to provide explanations for less complex parts.	Relevant for black box models. Example: ''Post-Hoc'' such as grad-CAM & transparent ''Protoypes'' based neural networks.
Semi-Local	Combines local and global explanations to provide insights on individual predictions and overall model characteristics.	Example: ''Prototypes'' and ''Concept'' based networks.

Table 4
Background distribution of participants in the study group.

Stakeholder group	Number of participants
Ethicists	2
Engineers	2
Retailers	2
Privacy experts	2
Members of society at large	2
Public decision-makers (City level, EU and US)	2

- **Goodness and Satisfaction:** assess the user's overall satisfaction and the quality of the explanations provided by the AI system. This metric measures whether the explanations help users understand how the system works, and whether the details provided are sufficient and actionable. Goodness and Satisfaction include whether the way the system works is understandable and whether the explanation increases confidence in the AI's results.
- **Curiosity:** is designed to determine whether the AI system can adequately address all potential questions a user might have about its operations and outcomes. This includes the system's ability to anticipate and answer questions about alternative decisions or actions it might have taken in different circumstances.
- **Trust:** assesses the user's confidence in the AI system. This metric is critical in determining whether users feel safe relying on the system, perceive its results as predictable, and believe it operates efficiently. A high level of trust is essential for the effective use of AI systems in sensitive or critical applications.
- **Performance:** focuses on the impact of explainability on user performance. This includes assessing whether explainability leads to improved user–system interaction outcomes and whether users feel that their understanding of the system positively affects their performance.

Each category addresses different aspects of the user's interaction with the AI system, helping us to evaluate and improve the system's explainability from multiple perspectives.

The methodology used was two focus groups composed of different stakeholders (ethicists, engineers, retailers, privacy experts, members of society at large, two public decision-makers at city level in the EU and the US) and the election of a panel of 10 members to vote by consensus on the questions and the scoring used. The questions were revised and perfected after consultation with a pool of experts. The scores were calculated as an average of the individual scores of each panel member. This approach ensured a balanced assessment and minimized individual bias. Where there were significant differences in scores between panel members, a consensus discussion was held to agree a final score (Pokholkova et al., 2024). Each focus group member was provided with a metric from 0 to 10 expressing the increasing severity of each question, as well as a handbook explaining the ethical issue/risk at stake in the context being considered (with a blank space for qualitative observations). To illustrate the composition of the study group, Table 4 gives a distribution of the participants:

Participants were selected with the intention of representing a broad range of stakeholders to ensure that multiple aspects of ethical considerations were addressed. This diversity is crucial for obtaining reliable results. Our approach is consistent with the methodology outlined by Pokholkova et al. (2024), which demonstrates that small, well-chosen groups can provide meaningful and reliable insights if members represent a diverse range of perspectives. The careful selection of diverse stakeholders and the structured, consensus-based methodology provide a strong foundation for reliable and meaningful results.

To maintain homogeneity and avoid background bias, we used the same weight for each member in the average rating elaboration. A detailed overview of these metrics and the specific aspects they cover are reported in Table 5. This Table provides a structured overview of each metric along with a scale for quantification, allowing for a systematic evaluation of the explainability of our AI systems.

5. Results and discussions

In this study, we use the VRAI deep learning framework, originally introduced by Paolanti et al. (2020) to evaluate our approach, which has been instrumental in advancing retail analytics applications. The framework uses three convolutional neural networks (CNNs) to perform simultaneous people counting, top view re-identification, and shopper-shelf interaction measurement in a single RGB-D video frame at a rate of 10 frames per second. This setup aims to provide comprehensive insights into shopper behaviour and store dynamics, addressing the needs of the modern retail environment where understanding consumer interactions can significantly influence management and marketing strategies. The VRAI framework was deployed in a retail store of approximately 1500 square metres, using 24 RGB-D cameras strategically placed to maximize coverage without overlap. This included two cameras at the store entrances and 22 cameras facing the shelves. Over a period of two years, this system collected data from five global locations, including Italy, China, Indonesia and the USA, creating a significant dataset of interactions from 10.4 million shoppers. However, while the system performed exceptionally well in terms of technical accuracy and reliability, the complexity of managing and interpreting vast amounts of data poses challenges, particularly in terms of scalability and real-time data processing. In addition, the assessment of bias and explainability highlights the need for continuous refinement of AI systems to ensure they remain fair, understandable and effective in different deployment contexts.

Following the biases identified in our AI Ethics Framework for Bias Detection (Tables 1 and 2), several biases were evaluated:

- **Population targeting bias:** There is a risk that the data collected and the algorithms used may not adequately represent the diversity of the global shopper population, potentially leading to biased analysis that unfairly favours certain demographics instead of others.
- **Socio-behavioural Context Bias:** The system's failure to account for different behavioural contexts across different cultures and store formats can lead to inaccuracies that limit the effectiveness of the insights generated while producing discrimination.
- **Historical bias:** Relying on historical data without continuous updates can perpetuate existing or historical socially-embedded stereotypes and behaviours, further compromising the system's accuracy and fairness.

The VRAI framework operates as a "black box" where the decision-making processes are not transparent, making it difficult for users to understand how conclusions are derived. This lack of accountability is particularly problematic in environments where understanding consumer behaviour patterns is critical to making not only strategic but also responsible business decisions. The evaluation of the VRAI framework against established explainability metrics is systematically summarized in Table 6, which provides a detailed breakdown of the scores across different categories such as "Goodness and Satisfaction", "Curiosity", "Trust" and "Performance". The framework shows limited ability to provide comprehensive and detailed explanations, with particularly low scores for helping users to understand how the system works and for completeness of information provided (scores: 1–3). The actionability of the explanations and their ability to convey the reliability and trustworthiness of the system are also critically low (scores: 2–3), highlighting significant gaps in current implementation. Although the framework performs relatively better in terms of efficiency and predictability (scores of 5–6), these attributes do not fully address the deficiencies in explainability that are critical to user confidence and effective use of the system. This categorical evaluation underlines the urgent need to improve the explainability features of the VRAI framework to ensure that it meets the ethical standards required for trustworthy AI applications in retail environments.

Table 5
Explainability metrics.

Category	Indicator	Scale (1–10)
(a) Goodness and satisfaction		
	Does the explanation help the user to understand how the system works?	1–10
	Is the explanation of how the system work satisfying?	1–10
	Is the explanation of the system sufficiently detailed?	1–10
	Is the explanation of how the system works sufficiently complete?	1–10
	Is the explanation actionable (i.e., it helps the user to know how to handle the system)?	1–10
	Does the explanation let the user know how accurate or reliable the system is?	1–10
	Does the explanation let the user know how trustworthy the system is?	1–10
(b) Curiosity		
	The user wants to know what the system did	1–10
	The user wants to understand what the AI system will do next	1–10
	The user wants to know why the AI system did not make some other decision	1–10
	The user wants to know what the AI system would have done if something had been different	1–10
(c) Trust		
	The user is confident the AI system works well	1–10
	The outputs of the AI system are very predictable	1–10
	The AI system is very reliable. The users can count on it to be correct all the time	1–10
	The user feels safe when they rely on the system	1–10
	The AI system is efficient	1–10
	The user is wary of the AI system	1–10
	The AI system can perform the task better than a novice human user	1–10
(d) Performance		
	User performance will improve as a result of being given satisfying explanations	1–10
	User performance may be affected by their level of epistemic trust	1–10
	User performance will be appropriate if the user has been able to explore the competence enveloped of the AI system	1–10

Table 6
Evaluation of the VRAI framework based on explainability metrics.

Category	Mean score (1–10)
Goodness and satisfaction	2.0
Curiosity	2.0
Trust	3.7
Performance	2.3

6. Enhancing trustworthiness in AI systems for consumer's behaviour understanding

In response to the challenges identified in the VRAI framework, particularly around explainability and the presence of bias, we propose several strategic enhancements aimed at strengthening the trustworthiness of the system. These improvements are designed not only to address the immediate gaps, but also to set a standard for future developments in AI-driven retail analytics systems.

As mentioned above, the tasks involved in the VRAI framework are people counting, classification for shopper interaction analysis and re-identification and semantic segmentation for people counting. In applications where privacy is paramount and regulations prevent the storage of images or videos, the implementation of post-hoc explainability methods becomes infeasible. Post-hoc methods typically require access to stored data to analyse how decisions were made after the fact, which contradicts privacy-first approaches necessary in sensitive environments. This limitation necessitates the adoption of transparent models like the Semantic Prototype Analysis Network (SPANet) (Wan et al., 2024), which are designed to offer real-time explanations. The SPANet is an interpretable object recognition method that enhances the clarity and comprehensibility of decision-making processes for users. It achieves this by simultaneously "highlighting the areas to focus on" and "explaining the reasons behind these focal points". Unlike other methods that apply concepts across the entire image, SPANet specifically aligns these concepts with localized areas, thereby assigning semantic labels to the identified part prototypes. This targeted approach helps make the interpretation of AI decisions more intuitive and contextually relevant. SPANet integrates seamlessly into systems where instant interpretation of AI decisions is critical, without the need for data retention. By embedding explainability directly into the operational process, SPANet provides immediate, understandable insights into the model's reasoning processes. This approach not only adheres to strict privacy requirements by eliminating the need for data storage but also enhances user trust and acceptance by clarifying AI actions as they occur. Thus, transparent models like SPANet are indispensable in scenarios where upholding privacy and providing clarity in AI operations are equally critical. SPANet offers a novel approach to real-time explainability by using both visual and textual elements. This is vital in a retail context, where interactions like product handling or shelf browsing need to be immediately understood and contextualized. SPANet utilizes part prototypes and semantic concepts to generate comprehensive, real-time explanations that are intuitively aligned with human reasoning. Prototypes such as "fingers" or "boxes" are visually identified in real-time, providing immediate clues about the nature of the interaction. Accompanying each visual prototype, semantic tags such as "taking product" or "returning product" offer contextual explanations that help staff understand shopper behaviour on a nuanced level. This integration of prototypes and concepts mirrors natural human explanatory processes, where we point out specific features and provide a narrative to explain what we observe (e.g. "that interaction was positive because customer was holding that product in his hand"). In the re-identification module, where top-view cameras capture shopper movements without storing any imagery to ensure privacy, SPANet plays a crucial role in providing understandable and immediate explanations for shopper re-identification. By employing SPANet, we ensure that every identification made by the system is accompanied by a semantic explanation, such as recognizing the same customer because "he's wearing a red hat and is 180 cm tall". This method not only maintains privacy but also enhances the trustworthiness and utility of the system by making AI decisions transparent and immediately understandable. Extending the existing VRAI framework to incorporate explainability within the people counting system, primarily achieved through semantic segmentation, marks a significant enhancement in understanding the underlying decisions made by our model. Traditional segmentation methods, while effective for class identification and boundary delineation, often lack transparency in their decision-making processes. To address this, we have integrated a novel explainability approach inspired by advancements in

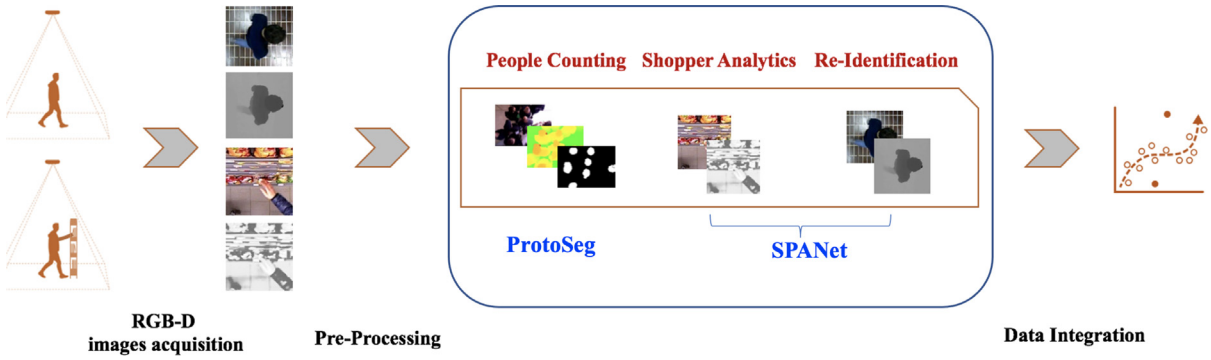


Fig. 2. Explainable VRAI framework for consumer behaviour understanding. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

interpretability techniques specifically tailored for image segmentation. State of art methods for semantic segmentation explainability include Seg-XRes-CAM (Hasany et al., 2023), an advanced version of the previously established Seg-Grad-CAM (Vinogradova et al., 2020). Seg-Grad-CAM, an extention of Grad-CAM (Selvaraju et al., 2017) is adept at explaining the entire segmentation map for a target class but falls short when detailed, localized explanations within the segmentation map are necessary. The Seg-XRes-CAM methodology, inspired by the capabilities of HiResCAM (Draelos and Carin, 2020), integrates spatial awareness into the explanation process, thereby allowing for more precise and relevant insights. Although relevant, these kind of methodology falls into the post-hoc categories and cannot be applied in real-time privacy-preserving systems like VRAI Framework, a different approach is indeed necessary. ProtoSeg (Sacha et al., 2023) introduces an interpretable method for semantic segmentation, distinguished by its use of prototypical parts. Unlike conventional segmentation approaches, which typically provide only class probabilities for each pixel, ProtoSeg leverages learned prototypes for each class to facilitate and clarify segmentation. This method involves using patches (cases) from the training set that correspond to specific parts of the segmented objects. For example, in the segmentation of a bus, ProtoSeg utilizes prototypes that might represent distinct features such as windows or wheels, highlighted in red and orange respectively. This approach not only generates segmentation but also provides a meaningful interpretation of the segmentation results by directly relating image areas to identifiable object parts. By applying Protoseg to the people counting segment, the model not only identifies and counts individuals but also provides localized explanations of the segmentation decisions before images are fed to the counting algorithm. Fig. 2 schematically shows the explainable VRAI framework that integrates AI ethics.

7. Conclusions and future works

This paper has extensively analysed the use of AI in the retail sector, with a focus on ensuring ethical, human-centred and trustworthy deployment. By critically examining both the potential and pitfalls of AI in retail, particularly in terms of privacy, surveillance and biased outcomes, we have highlighted the urgent need for ethical considerations to be embedded in AI systems. By bridging the gap between theoretical AI ethics frameworks and practical retail applications, this paper contributes to a more nuanced understanding of what it means to design and use AI responsibly in retail environments. As AI continues to reshape the retail landscape, ensuring that these technologies are developed and implemented in an ethical manner remains a primary concern. Our work aims to serve as a foundational guide for stakeholders, helping them navigate the complex ethical landscape while fostering the development of AI systems that are not only technologically advanced but also socially responsible and trustworthy. The introduction of the VRAI framework has been crucial in demonstrating the practical applications of these technologies but has also highlighted the complexities involved in achieving transparency and fairness. Our approach merged high-level AI ethics principles with the specific requirements of the retail context to provide a unique, context-sensitive ethical framework. This framework not only addresses the technical aspects of AI in retail but also aligns with broader ethical and legal standards, particularly in the European context.

Looking ahead, several areas need further exploration to improve the implementation and effectiveness of ethical AI systems in retail. Continuous improvement in the explainability of AI systems such as VRAI is critical. We aim to conduct comparative evaluations of the VRAI framework with other state-of-the-art human behaviour analysis approaches to further investigate the performance and robustness of the framework. This comparative analysis will help to consolidate the VRAI framework's status in the field and provide valuable insights into best practices for implementing trustworthy AI systems in different retail environments. We also plan to incorporate larger and more diverse samples to enhance the generalizability of our findings and continuously update our AI ethical framework to address new challenges and opportunities in the evolving retail environment, ensuring the development of trustworthy AI systems that benefit all stakeholders. Future research should explore advanced methodologies that can provide clearer insights into AI decision-making processes in real time. As biases in AI can lead to significant ethical and operational risks, future efforts must focus on developing more sophisticated techniques to identify and mitigate biases at every stage of AI system development and deployment. To ensure that ethical frameworks remain relevant, they need to evolve with technological advances and changes in consumer behaviour. This includes regular updates based on new research and stakeholder feedback. Active participation in shaping regulatory frameworks is necessary to ensure that ethical considerations in AI keep pace with global standards. This includes contributing to policy discussions and compliance measures that govern the use of AI in different regions. Implementing additional real-world case studies, such as expanded applications of the VRAI framework, will help validate the proposed ethical guidelines and demonstrate their practical benefits and limitations.

CRediT authorship contribution statement

Simona Tiribelli: Methodology, Conceptualization. **Benedetta Giovanola:** Methodology, Conceptualization. **Rocco Pietrini:** Writing – original draft, Validation, Investigation, Formal analysis, Data curation. **Emanuele Frontoni:** Writing – review & editing, Supervision. **Marina Paolanti:** Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This work is funded by Egocentric and exocentric views for an object-level human behaviour analysis and understanding through tracking in complex spaces (EXTRA EYE) project, Piano Nazionale di Ripresa e Resilienza Missione 4 - Componente 2 – Investimento 1.1 "Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)", Codice CUP D53D23008900001.

References

AI, H., 2019. High-level expert group on artificial intelligence. In: Ethics Guidelines for Trustworthy AI. 6, European Commission.

Anica-Popa, I., Anica-Popa, L., Rădulescu, C., Vrîncianu, M., 2021. The integration of artificial intelligence in retail: benefits, challenges and a dedicated conceptual framework. Amfiteatru Econ. 23 (56), 120–136.

Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Schneeberger, D., Stoeger, K., Holzinger, A., 2023. Quod erat demonstrandum?-towards a typology of the concept of explanation for the design of explainable AI. Expert Syst. Appl. 213, 118888.

Contreras, B.P., Hoffmann, A.N., Slocum, T.A., 2021. Ethical behavior analysis: Evidence-based practice as a framework for ethical decision making. Behav. Anal. Pract. 1–16.

Corrêa, N.K., Galvão, C., Santos, J.W., Del Pino, C., Pinto, E.P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., et al., 2023. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. Patterns 4 (10).

Ding, W., Abdel-Basset, M., Hawash, H., Ali, A.M., 2022. Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. Inform. Sci. 615, 238–292.

Draelos, R.L., Carin, L., 2020. Use HiResCAM instead of grad-CAM for faithful explanations of convolutional neural networks. arXiv preprint arXiv:2011.08891.

Ferracuti, N., Norscini, C., Frontoni, E., Gabellini, P., Paolanti, M., Placidi, V., 2019. A business application of RTLS technology in intelligent retail environment: Defining the shopper's preferred path and its segmentation. J. Retail. Consum. Serv. 47, 184–194.

Fildes, R., Ma, S., Kolassa, S., 2022. Retail forecasting: Research and practice. Int. J. Forecast. 38 (4), 1283–1318.

Floridi, L., 2014. The Fourth Revolution: How the Infosphere Is Reshaping Human Reality. OUP Oxford.

Frontoni, E., Paolanti, M., Pietrini, R., 2019. People counting in crowded environment and re-identification. RGB-D Image Anal. Process. 397–425.

Giovanola, B., Tiribelli, S., 2022. Weapons of moral construction? On the value of fairness in algorithmic decision-making. Ethics Inf. Technol. 24 (1), 3.

Giovanola, B., Tiribelli, S., 2023. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. AI & Soc. 38 (2), 549–563.

Hasany, S.N., Petitjean, C., Mériaudeau, F., 2023. Seg-xres-cam: Explaining spatially local regions in image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3732–3737.

Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J., 2018. Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.

Jobin, A., Ienca, M., Vayena, E., 2019. The global landscape of AI ethics guidelines. Nat. Mach. Intell. 1 (9), 389–399.

Kelly, E.M., Greeny, K., Rosenberg, N., Schwartz, I., 2021. When rules are not enough: Developing principles to guide ethical conduct. Behav. Anal. Pract. 14, 491–498.

Kliestik, T., Kovalova, E., Lăzăroiu, G., 2022. Cognitive decision-making algorithms in data-driven retail intelligence: consumer sentiments, choices, and shopping behaviors. J. Self-Govern. Manage. Econom. 10 (1), 30–42.

Martini, M., Paolanti, M., Frontoni, E., 2020. Open-world person re-identification with rgbd camera in top-view configuration for retail applications. IEEE Access 8, 67756–67765.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Comput. Surv. (CSUR) 54 (6), 1–35.

Migliorelli, L., Tiribelli, S., Cacciatore, A., Giovanola, B., Frontoni, E., Moccia, S., 2023. Accountable deep-learning-based vision systems for preterm infant monitoring. Computer 56 (5), 84–93.

Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E., 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. Front. Big Data 2, 13.

Paolanti, M., Pietrini, R., Mancini, A., Frontoni, E., Zingaretti, P., 2020. Deep understanding of shopper behaviours and interactions using RGB-D vision. Mach. Vis. Appl. 31, 1–21.

Pascucci, F., Nardi, L., Marinelli, L., Paolanti, M., Frontoni, E., Gregori, G.L., 2022. Combining sell-out data with shopper behaviour data for category performance measurement: The role of category conversion power. J. Retail. Consum. Serv. 65, 102880.

Pokholkova, M., Boch, A., Hohma, E., Lütge, C., 2024. Measuring adherence to AI ethics: a methodology for assessing adherence to ethical principles in the use case of AI-enabled credit scoring application. AI Ethics 1–23.

Rossi, L., Paolanti, M., Pierdicca, R., Frontoni, E., 2021. Human trajectory prediction and generation using LSTM models and GANs. Pattern Recognit. 120, 108136.

Sacha, M., Rymarczyk, D., Struski, Ł., Tabor, J., Zieliński, B., 2023. Protoseg: Interpretable semantic segmentation with prototypical parts. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1481–1492.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Slocum, T.A., Detrich, R., Wilczynski, S.M., Spencer, T.D., Lewis, T., Wolfe, K., 2014. The evidence-based practice of applied behavior analysis. Behav. Anal. 37, 41–56.

Suarez, V.D., Marya, V., Weiss, M.J., Cox, D., 2023. Examination of ethical decision-making models across disciplines: Common elements and application to the field of behavior analysis. Behav. Anal. Pract. 16 (3), 657–671.

Suresh, H., Guttag, J., 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. pp. 1–9.

Tiribelli, S., 2024. Who decides what online and beyond: freedom of choice in predictive machine-learning algorithms. In: Ethics in Online AI-Based Systems. Elsevier, pp. 299–321.

Vinogradova, K., Dibrov, A., Myers, E.W., 2020. Towards interpretable semantic segmentation via gradient-weighted class activation mapping. In: Proceedings of the AAAI Conference on Artificial Intelligence. http://dx.doi.org/10.1609/aaai.v34i10.7244.

Wan, Q., Wang, R., Chen, X., 2024. Interpretable object recognition by semantic prototype analysis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 800–809.

Wilkenfeld, D.A., McCarthy, A.M., 2020. Ethical concerns with applied behavior analysis for autism spectrum "disorder". Kennedy Inst. Ethics J. 30 (1), 31–69.