Document Title	**Educators' guide for promoting media-related algorithmic transparency**
Project Title and acronym	Cyprus Center for Algorithmic Transparency (CyCAT)
H2020-WIDESPREAD-05-2017-Twinning	Grant Agreement number: 810105 — CyCAT
Deliverable No.	D4.2
Work package No.	WP4
Work package title	Promoting algorithmic transparency
Authors (Name and Partner Institution)	Miranda Christou (OUC) Michalinos Zembylas (OUC)
Contributors **(Name and Partner Institution)**	Maria Kasinidou (OUC) Paraskevi Kleanthous (OUC) Jahna Otterbacher (OUC)
Reviewers	Jo Bates (USFD) Nandu Chandran Nair (UNITN)
Status **(D: draft; RD: revised draft; F: final)**	F
File Name	D4.2_Educators_Guide_M18
Date	31 March 2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 810105.

Draft Versions - History of Document				
Version	**Date**	**Authors / contributors**	**e-mail address**	**Notes / changes**
v1.0	15/9/19	J. Otterbacher	jahna.otterbacher@ouc.ac.cy	Initial version
v2.0	25/9/19	M. Zembylas	m.zembylas@ouc.ac.cy	Addition of rationale
v3.0	20/2/20	M. Kasinidou	maria.kasinidou@ouc.ac.cy	Simplification of case studies; addition of lesson
v4.0	27/2/20	M. Kasinidou	maria.kasinidou@ouc.ac.cy	Addition of translation

Abstract

Deliverable D4.2 is an easy-to-read guide for educators, summarizing the core technical concepts surrounding algorithmic system transparency and explaining how these can be taught to secondary school students. In particular, this document "translates" the framework for end-to-end fairness management in algorithmic systems, presented in D4.1. It explains to teachers the importance of raising students' awareness of algorithmic processes and algorithmic bias, by contextualizing the CyCAT framework within familiar pedagogical principles. Finally, this document provides example lesson plans (developing the objectives, materials, activities, and finally, the evaluation) enabling teachers to implement them in their own classrooms.

Keyword(s):	Algorithmic bias, core concepts, Cyprus public schools, educators' guide, media-related algorithmic transparency

Contents

1. Executive Summary

D4.2 presents an easy-to-use guide for secondary school teachers, focused on explaining - in everyday language - the problem of social and cultural bias in algorithmic systems as well as ways in which such biases can be mitigated via *fairness management* processes. In D4.1, a framework for end-to-end fairness management in algorithmic systems was presented, based on an extensive review of the state-of-the-art (i.e., the literature review presented in D3.1). Of key importance in D4.1 is the inventory of technical concepts surrounding algorithmic systems and fairness management, which can and should be understood by each of the user groups addressed.

Therefore, following the work of D4.1, the goal of D4.2 is to "translate" the parts of the framework and the relevant concepts, in a manner that is meaningful and useful for one particular user group - secondary school teachers. Specifically, this guide is written with a particular audience in mind - teachers working in the public school system in Cyprus. As such, the guidelines serve as the basis for the development of the teacher training, which will be carried out in WP5 (deliverables D5.1, D5.2). The Greek translation of this guide, which shall be used in the context of the training, is presented in the Annex.

2. Introduction: Purpose of the Guidelines

This Guide aims to help you promote *algorithmic literacy* in your classroom. Algorithmic processes that are proprietary and technically complex now play a key role in the information access technologies we use on a daily basis. Whether we are searching for information on the Web, browsing NetFlix for a movie to watch, or socializing on a social media platform, algorithmic processes mediate nearly all of our access to information. Research has demonstrated that many users do not realize the presence of these algorithms, and believe that they have access to all the information that exists in a given platform/context. Similarly, we often overlook the fact that technologies are not neutral; they embed and reflect human values (e.g., in a search engine, efficiency and speed, but also a preference for particular sources of information), and can thus shape our views and practices. For instance, a Google search for the keyword "nurse" reflects the human beliefs that most nurses are women.

For all of these reasons, it is important to be aware of the role that algorithms play and how they may impact citizens' abilities to stay informed and engaged. This Guide will help you achieve an understanding of the nature of algorithmic information access systems and how social biases may arise within such systems. Along the way, we will present concrete examples of real cases within systems that you likely use regularly. After that, we shall review the pedagogical principles that guide our approach to fostering algorithmic literacy. Finally, we shall present examples of interactive, classroom activities that can be used to raise students' awareness of the role of algorithms in information access. During the face-to-face seminars that shall accompany this Guide, you will be encouraged to tailor these activities, in formulating a lesson plan that will be effective in light of your particular students' needs.

3. Definitions and Core Technical Concepts

3.1 Algorithmic Processes and Systems

An algorithm, according to the Merriam-Webster dictionary, is "a step-by-step procedure for solving a problem or accomplishing some end." It is a means to organize the thoughts and actions involved in working

towards a desired outcome. A simple example is a recipe, which dictates to the cook not only the list of ingredients and utensils necessary to make the desired food, but also, the steps that must be followed and the appropriate order. Thus, a recipe is a *fully transparent* algorithm, which clearly outlines the inputs (i.e., ingredients, utensils), the processing (i.e., steps that must be followed as well as any assumptions or things of which the cook should be careful), as well as the expected output (i.e., the dish being prepared, the number of servings, etc.)

Other examples of algorithms in our "offline lives" include sorting a set of paper files by the date they were produced, or assembling a piece of furniture that one has purchased (e.g., an item from IKEA, which characteristically consists of many different parts and is sold with a "how-to" assembly guide). In each of these cases, to reach the desired outcome (i.e., a stack of files that are in order by date; a piece of furniture that is assembled and fully functional), one needs to know the inputs to the process, as well as the step-by-step process that will lead us to the outcome.

Information systems rely extensively on digitally implemented algorithms (i.e., encoded in a computer language) to carry out operations leading to a desired functionality and/or end result for their users. For instance, in a company payroll system that computes employees' paychecks, an algorithm specifies all of the detailed instructions necessary to accomplish this complex task. First, the system requires all of the appropriate input data for a given employee (e.g., her rate of pay and overtime, the number of hours worked in the given time period, as well as any necessary personal information required in the respective country, such as marital status and/or income tax bracket). Next, it must specify all of the steps necessary in calculating the gross salary (e.g., multiply the rate by the time worked, multiple the overtime rate by any overtime hours worked, and sum these), any deductions the employer must make (e.g., taxes, social insurance, etc.) and finally, the net salary that should be paid out to the employee.

3.2 Algorithmic Systems that Learn

The above examples constitute algorithmic processes that are *static*; in a process such as a recipe, or a system for employee payroll, the algorithm is transparent and does not change automatically.[1] In contrast, the algorithms behind the modern interactive systems we use today to access information on the Web and social media, tend to be very dynamic. That is, they are designed in a manner that allows them to constantly learn and improve their performance, based on data that reflects users' behaviors and preferences when using the system.

Figure 1 depicts the basic architecture of an algorithmic system that learns from data (i.e., is based on *machine learning* techniques). As observed, there are five macro components to such a system. These components are explained below, using the example of a *Web search engine* (e.g., Google search or Microsoft Bing).

[1] Of course, there are circumstances under which changes in the algorithm will be necessary. In order to be successfully executed at a high altitude, a recipe for a cake will need readjustment. In a payroll system, changes in taxation law will require the system's algorithm to be updated. However, we assume that these changes are made manually; they are not based on data-driven insights and do not happen automatically.

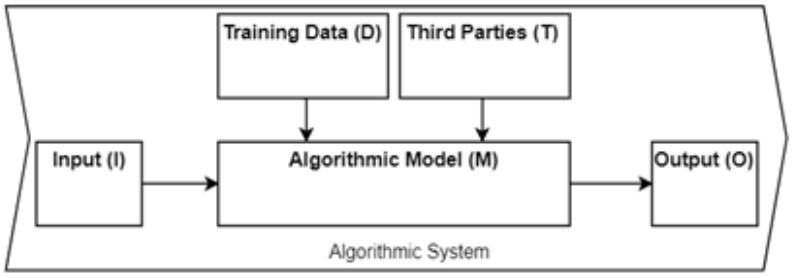


Figure 1. Basic architecture of an algorithmic system.

- **Input (I)**: When using the system, the user inputs some particular value(s) in order to run a given instance of the system. In a Web search, the user provides a set of keywords, which express her need for information on a given topic.[2]
- **Output (O):** The output is the information produced by the system, in response to the user's input. In the Web search case, this is the ranked set of Web pages that the algorithms have deemed to be the most likely to be useful for the user.
- **Algorithm (M):** The algorithmic model of the system is its core. This is the system component that, after having been trained from the data, maps a given input to a given output. In the case of a modern, proprietary search engine, the Algorithmic Model is actually not a single algorithm, but rather, an entire set of algorithmic processes. However, perhaps the most frequently discussed algorithm used in a search engine is the *ranking algorithm*, which assigns a score to each Web page retrieved in response to the user's keywords, such that the pages can then be presented to the user in ranked order.
- **Training Data (D):** The training data is used to train the Algorithmic Model (M) when some machine learning techniques are applied. As mentioned, in the case of information access systems such as search engines, the training data consists of observations that concern users' behaviors, personal attributes and information preferences. As a specific example in the case of Web search, a training dataset for the purpose of developing a ranking algorithm, would contain information about the types of Web pages that users find to be relevant to a given topic.
- **Third Party Constraints (T):** In some cases of algorithmic systems, a third party (i.e., not the user of the system, nor its developer) imposes some constraints on the manner in which the system works. This could be the owners/operators of the system, regulators, and others that influence the use and outcomes of the system. An example in the case of Web search is when operators must constrain the system's behaviors in order to comply with local laws. For instance, search engines often constrain the search results (i.e., outputs) that it presents to users searching for information using a real person's name, to comply with European Union privacy regulations.[3]

3.3 Algorithmic Information Access Systems

Having examined the characteristics of a dynamic algorithmic system, we now provide some examples of information access systems with which you and your students likely interact regularly. In particular, three classes of information access (IA) systems are discussed below. First, they are described generally in terms

[2] It should be noted that more experienced users often input more than just keywords, using advanced configurations in order to specify a need for information from particular sources, in a given language, etc.
[3] The "Right to be Forgotten."

of their functionality as well as their components, using the terminology and abbreviations introduced in Figure 1. Next, specific examples of each class of system are provided.

3.3.1 Recommender Systems

These are algorithmic systems that provide specific suggestions to users during their interaction with the system. Targeted advertising (while using the Web or a mobile application) is a prime example of a recommendation system. Suggestions (e.g., for businesses that are geographically nearby a user viewing an app on a mobile device) are provided according to what the system knows about the user (i.e., her demographic attributes, behaviors and preferences) as well as contextual information (e.g., the geographic location, day of the week, time of day, current weather, etc.) However, in reality, the suggestions provided may or may not be the best choices for the users; such recommendations may also be sponsored by paid third parties (e.g., an advertiser that paid for their ads to be delivered to targeted groups of people).

Examples of recommendation systems:

- Online targeted advertising (e.g., via Google AdSense)
- YouTube video recommendations ("Up Next" videos)
- Netflix movie recommendations ("Popular on Netflix," "Recently Watched")
- Online shopping recommendations (e.g., Amazon's "Customers who viewed this item also viewed…").

In a recommender system, the input (I) consists of the behavioral cues of the user; in other words, what the user is currently viewing or doing in the system. The system output (O) is the set of recommendations provided to the user. The training data (D) consists of previous observations of users, which would allow the algorithmic model (M) to map the current user attributes (demographics, behaviors, location) to items likely to be of interest to her. Finally, third party constraints (T) might include not only those mentioned previously, but also additional information provided by other users (e.g., ratings on a given item, which could then influence the model's assessment of the item's quality or appropriateness).

An example of a recommender system is Amazon. Amazon displays recommendations (O) to the user based on the user's profile and order history (D). Figure 2 depicts a user's Amazon home page. The user recently purchased a photo frame from Amazon.co.uk (I). In Figure 2 can be seen the relevant recommendations (O) from Amazon to the user that include photo frames and other relevant products (Home decoration).

Another example of a recommender system is eBay. Figure 3 shows the recommendations (O) that the eBay algorithm (M) came up with for a user who recently was looking for a water bottle (I).

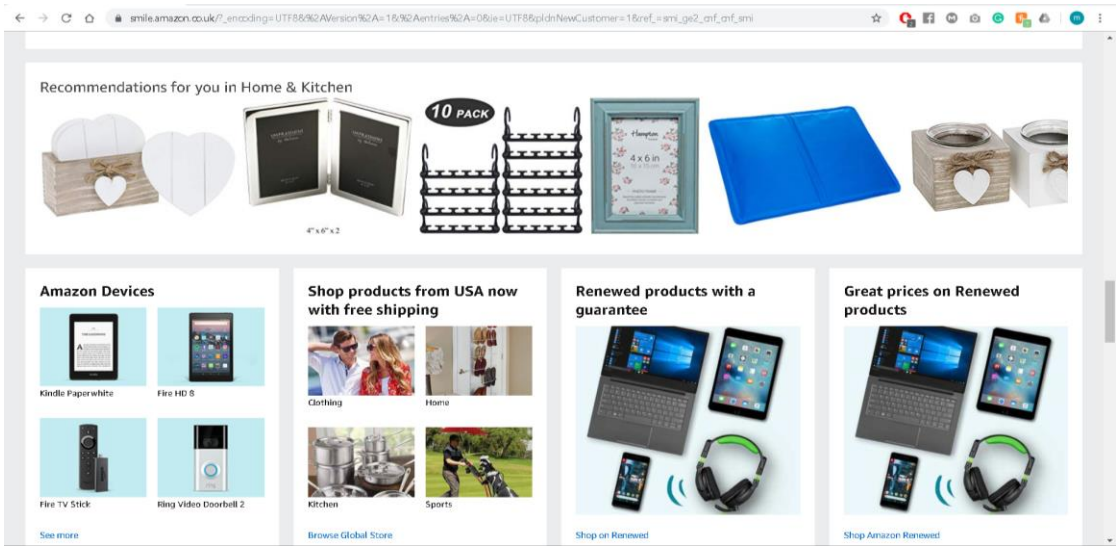


Figure 2: Amazon recommendations (O) based on user profile and previous purchases (I).

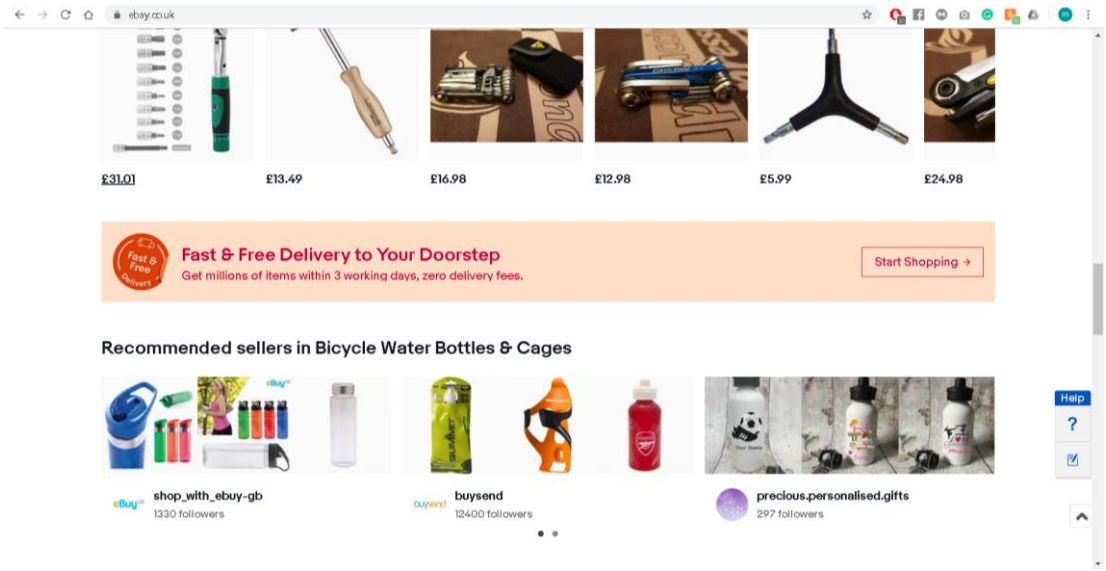


Figure 3: Ebay.co.uk recommendations (O) based on the user's search and profile (I).

A different recommender system is Booking.com. Booking.com's recommendations are not only based on user profile and her recent actions (I) but other contextual information such as the user's geographic location (I). A user who is currently located in Cyprus (I) is getting recommendations for cities nearby such as Limassol and Paphos (O) (Figure 4).

Netflix is another widely used recommender system. Netflix movie recommendations are based on the user's profile and the user's interaction with the system (D). Figure 5 depicts recommendations provided by Netflix to a user who recently has watched a Christmas movie (I). Netflix suggested to the user other Christmas movies (O). Also, Netflix suggested to the user, movies (Top picks) based on her profile and her interaction with the system (previously watched movies, search etc.).

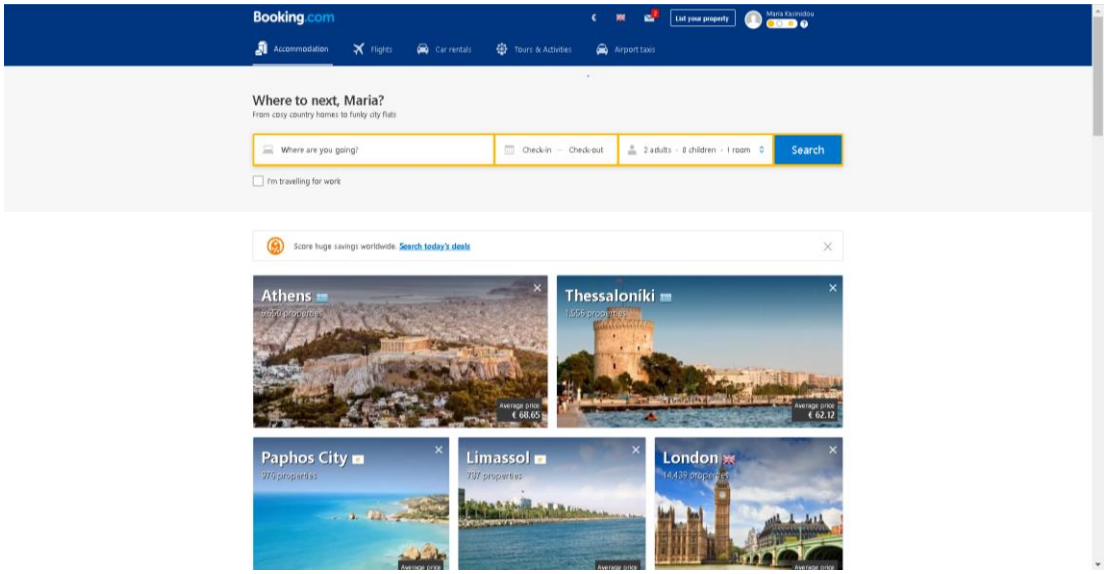


Figure 4: Booking.com recommendations (O) based on user's location and profile (I).

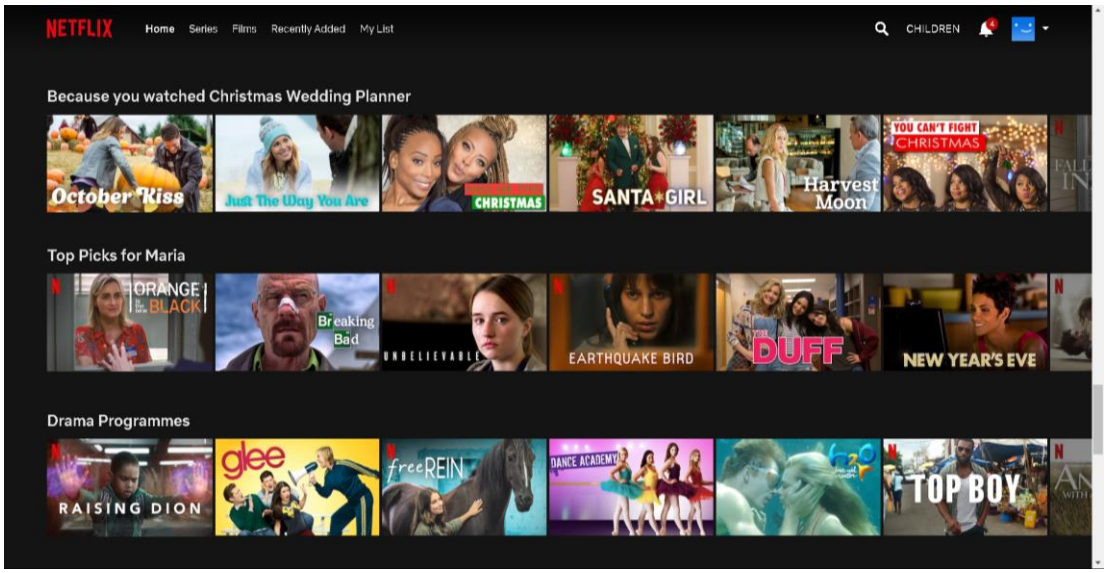


Figure 5: Netflix recommendations (O) based on previously watched movies and profile (I).

3.3.2 Search Engines

In contrast to recommender systems, within a search engine, the user explicitly expresses her need for information to the system in the form of a query (i.e., set of keywords and other optional parameters). This is the input (I) to the system, which in return, provides the user with a set of results (O) (i.e., Web pages, images, videos) that it predicts as being likely to satisfy the user's information needs. At the most basic level, the algorithmic model (M) learns from datasets (D) that indicate which types of items are relevant to which topics and keywords.

For instance, training datasets might consist of previous interactions within the system, thus recording which items users viewed after submitting a given information query. As previously mentioned, modern proprietary search engines consist of many algorithmic processes. Many of these are concerned with the localization of search results (e.g., prioritizing results that are geographically and/or culturally close to the

user) as well as personalization of results, based on what the system has learned about the user (e.g., by tracking her history of behavior during interactions with the system).

Finally, constraints on the system's behaviors may be imposed by third parties (T), particularly in order to comply with the law in various regions.

Examples of search engines:

- Google search (Web, images, video)
- Microsoft Bing
- DuckDuckGo ("The search engine that doesn't track you")
- Gibiru ("Uncensored, anonymous search")
- Yandex
- StartPage
- SwissCows ("The family friendly search engine")

Google is the most widely used search engine. An example of a Google search using the keywords "George Michael" (I) can be seen in Figure 6. Google returned a set of results including web pages, images, videos (O) which were the most relevant to the user's query.

Figure 7 shows a message (O) from Google interface to user after a query for "George Michael". Google detects user's location as "Strovolos" (I) and informs the user that it may have modified (T) the results of her search due to compliance with EU data protection laws.

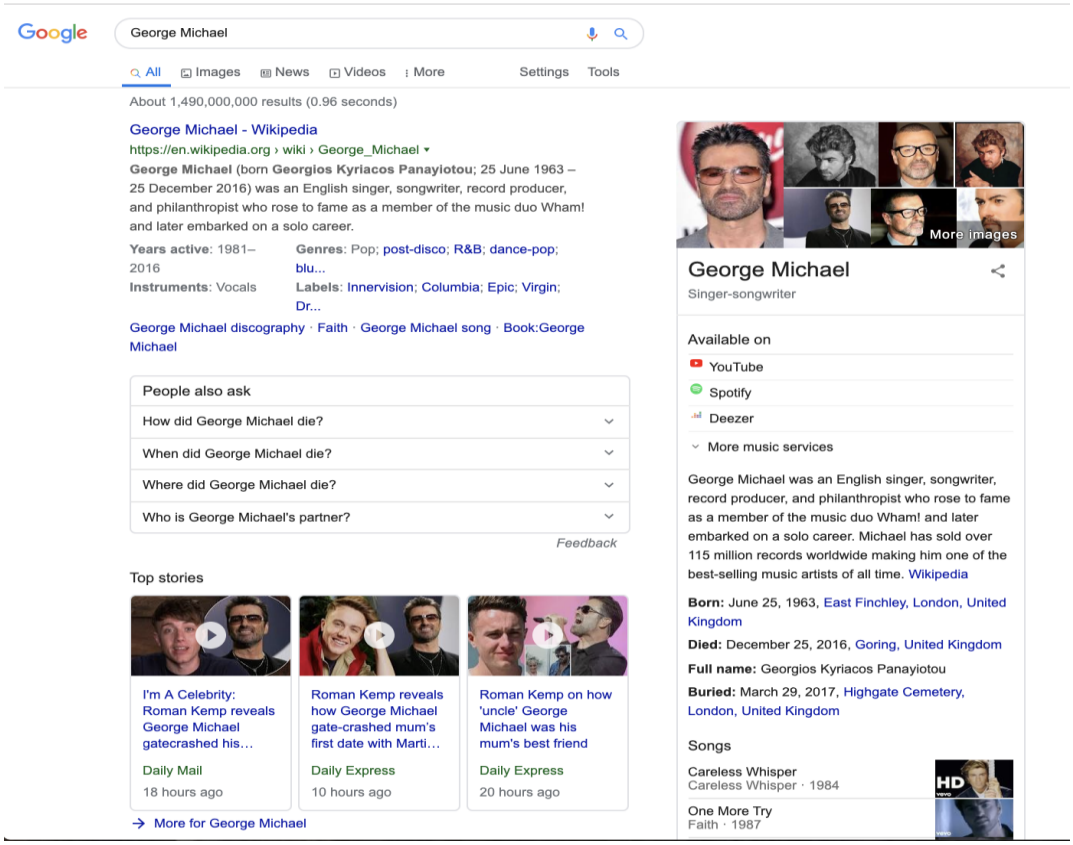


Figure 6: Google search results (O) for keywords "George Michael" (I).

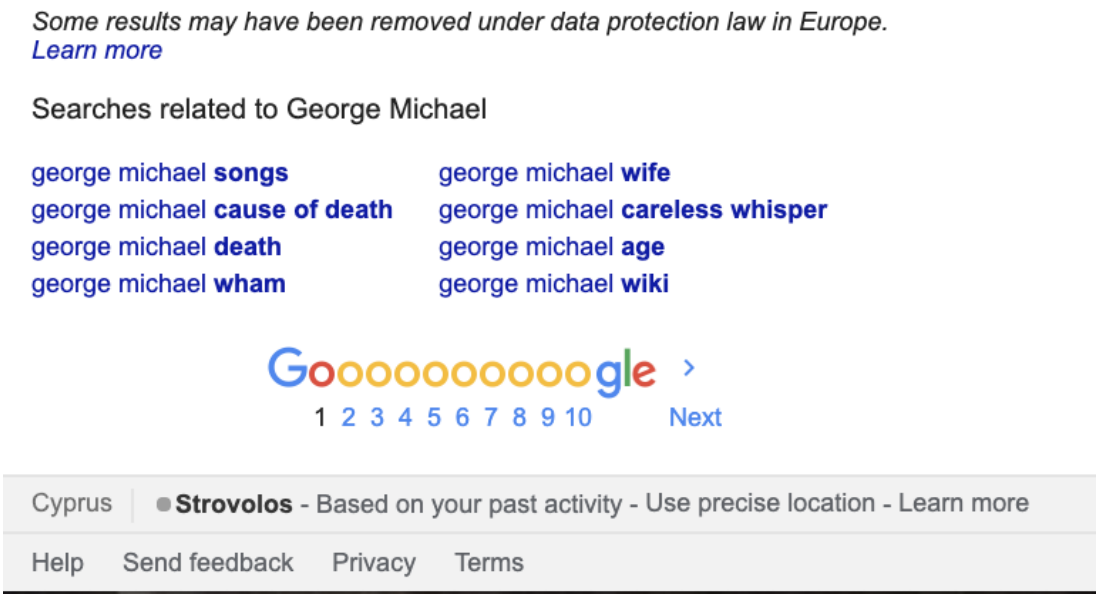


Figure 7: Google interface notes that the results (O) may have possible modifications (T) due to compliance with EU data protection laws; the user's location has been detected as being "Strovolos" (I).

3.3.3 Social Media News Feeds

Although many users may be unaware of it, the flow of posts within social media platforms are algorithmically curated. This is, of course, necessary due to the large volume of posts being made continuously on popular platforms such as Facebook, Twitter, Instagram, and LinkedIn. As social media platforms are proprietary, it is impossible to know exactly how their algorithmic curation processes work. However, we can examine the case of Facebook, and what it has disclosed to the public in the past year, following changes made to its News Feed after the Cambridge Analytica scandal.

According to a series of 2019 posts at the Facebook blog, the company historically tried to build algorithmic models (M) that predicted which content is most relevant to a user, in order to present an ordered feed of posts to her (O). "We've historically predicted who people might want to hear from based on signals like how often they interact with a given friend, how many mutual friends they have and whether they mark someone as a close friend." Such signals served as input (I) to the model. Training data (D) consisted of users' historical interactions on the platform. Facebook has mentioned the possible use of constraints (T) in order to reduce misinformation and political propaganda on the platform.

With respect to improvements in its algorithmic model, Facebook announced in May 2019, that "we have updated our algorithm to prioritize the Pages and groups we predict an individual may care about most. Some of the indicators of how meaningful a Page or group is might include how long someone has followed a Page or been a part of a group; how often someone engages with a Page or group; and how often a Page or group posts."

Another popular social media platform that, due to its large volume of posts, uses algorithmic procedures to select the 'appropriate' posts for each user is Twitter. CyCAT twitter account follows accounts related to education, research and other relevant accounts (I). Figure 8 depicts the CyCAT's News Feed page. The

posts which appear in the News Feed page are ordered based on the most relevant content to CyCAT's profile (O).

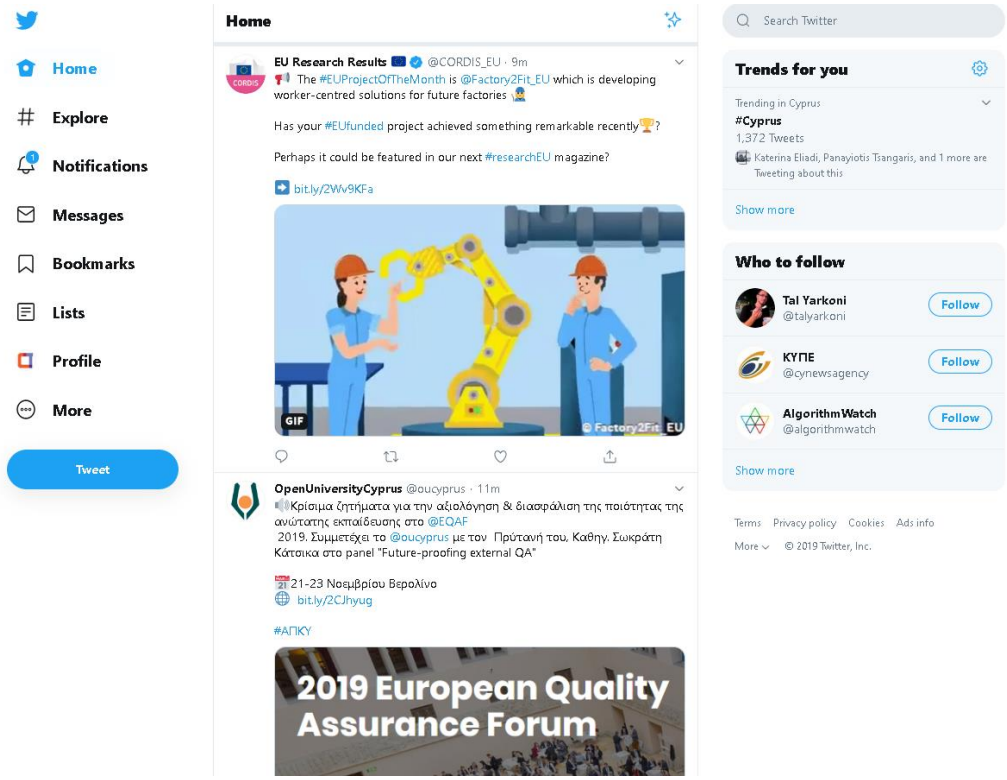


Figure 8: CyCAT's News Feed Twitter Page, posts order (O) presented is relevant (M) to user's behavior (I).

3.4 Biases in Algorithmic IA Systems and their Causes

Now, we turn our efforts toward examining the potential for algorithmic IA systems to exhibit social and cultural biases. First, we provide some examples of biases in the three classes of IA systems previously explored, which have been discussed in the Press. After that, we shall examine three causes of such biases in algorithmic systems: data bias, processing bias, and human bias.

3.4.1 Documented Biases in IA Systems

Here we discuss specific examples of algorithmic bias in IA systems, highlighting the following:

- **Fairness**: How were particular individuals or social groups discriminated against by the system?
- **Accountability**: Are there mechanisms by which the system (its owners and/or developers) can be held accountable for the observed unfairness?
- **Transparency**: Given that the systems below are proprietary, their algorithmic processes are protected by trade secrets; they are not transparent. Are there mechanisms within the system (and the user interface) that aim to **explain or interpret** the system's observed behaviors to the user.

Recommender Systems

An example of biases in recommender systems racist Google's ads which were more likely to recommend a criminal related to ad for black-sounding names. After an online search for "Latanya Farrell" (African American-sounding name) two ads were displayed as related to the search (Figure 9a). The first ad indicates

that she may have been arrested but there is no arrest record for her in the ad's link (instantcheckmate.com, shown in Figure 9b).

On the other hand, an online search for "Jill Foley" (white-sounding name in the American context) led to three neutral ads (Figure 9c), even if there is an arrest record in instantcheckmate.com for her (Figure 9d) (reflecting unfairness - racial discrimination). Google ads allow the reader to learn why a specific ad is displayed by clicking on the ⓘ icon in the ad banner. The icon is linked with a web page explaining why this ad appeared. However, the given explanation that reveals why an ad appeared is nothing more than a message that the ad matched the combination of the first and the last name searched from the user. There is no mechanism within the system which tries to explain/interpret the system's behaviors.

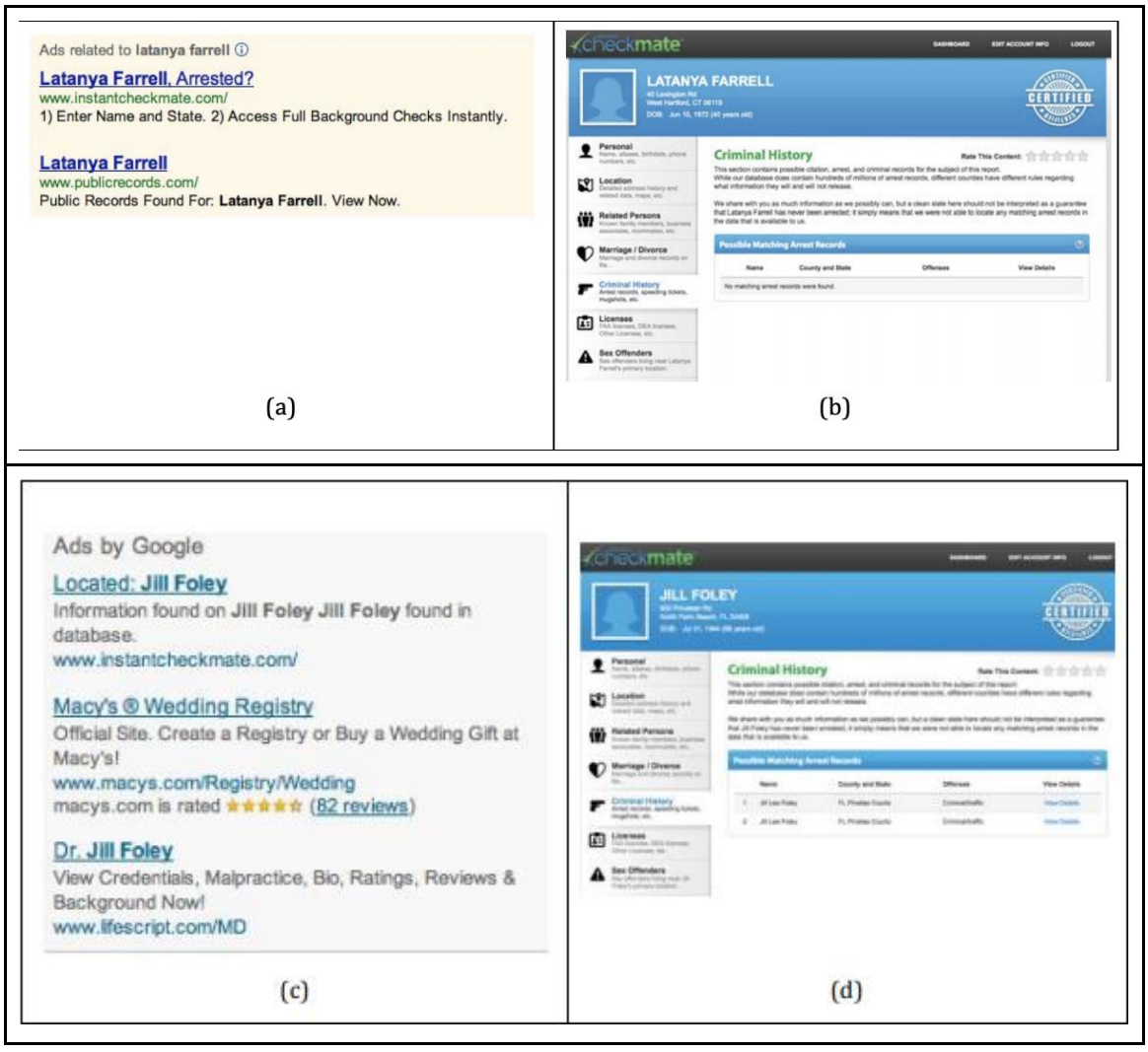


Figure 9: A 2013 study by L. Sweeney showed systematic racial bias in Google Ads.[4]

Another example of biases in recommender systems is Spotify's recommendation songs. Martina McBride (female singer) attempted to create a playlist for "Country Music" on Spotify. Spotify's recommendations included songs mostly from male singers for her playlist (reflecting unfairness - gender discrimination).

[4] Sweeney, L. (2013). Discrimination in online ad delivery. *arXiv preprint arXiv:1301.6822*.

She said she needed to refresh 14 times the recommendation page to get a song from a female singer (Figure 10). In this case, Spotify's recommendations were based on the title of the playlist. Nevertheless, Spotify does not provide more information about how/why the system chose the specific songs.

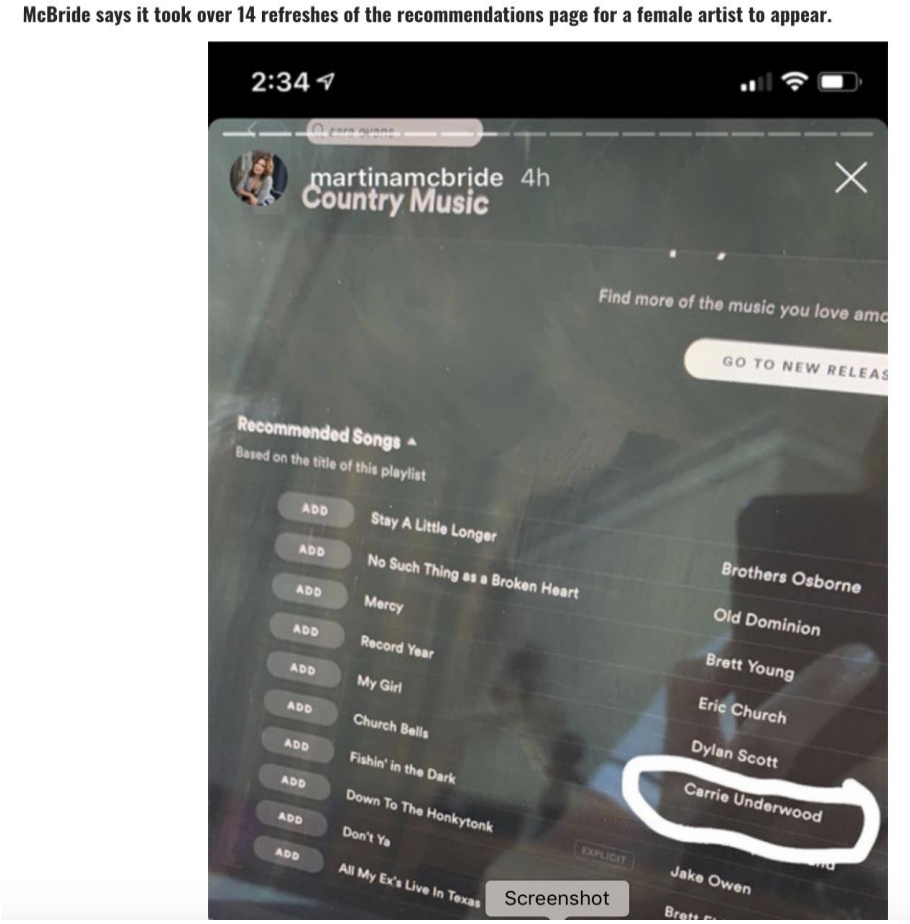


Figure 10: Are Spotify's recommendations sexist? Recordist artist Martina McBride and others have suggested the algorithm rarely suggests women artists to users.[5]

Search Engines

Examples of biases can be found in image search results provided by search engines such as Bing and Google. Both Bing and Google use algorithms to rank the results of a query. Both search engines use deep learning and try to understand the content of an image. However, how these algorithms work is a black box for the user and there is no mechanism to explain/interpret why a query's result specific images in a specific order. Results for a Bing search for the keyword "Nurse" are presented in Figure 11. As can be seen, the majority of the result images are depicting female nurses, which may be perceived as being unfair. Another example of a Bing search for the keywords "intelligent person" is shown in Figure 12. In this case the majority of the result images are displaying men (reflecting unfairness - gender discrimination).

[5] https://www.digitalmusicnews.com/2019/09/11/martina-mcbride-spotify-sexist/

Figure 11: Bing results for the keyword "Nurse"

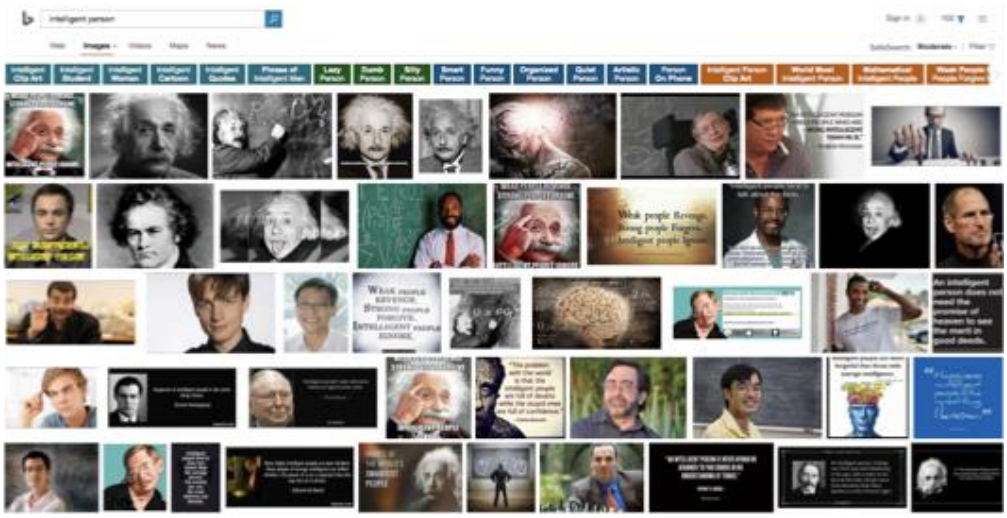


Figure 12: Bing results for the keywords "intelligent person"

Also, more examples of biased results in searches have been made known through social media. In 2016 a Twitter user shared a post about google search results which perceived "racism" and the post went viral. She posted the image results of a search for "unprofessional hairstyles for work" and "professional hairstyles for work". The results displayed black women's hair as "unprofessional" (Figure 13a), while the results for professional hairstyles for work presented mostly white women's hairstyles (reflecting unfairness - racial discrimination).

Another Twitter user posted a video of himself using Google searching for "three white teenagers" and "three black teenagers". The results of both searches can be seen in Figure 13. The first search produced images with smiling, happy white teenagers. In contrast, the second search produced "negative images" with images of mug shots (reflecting unfairness - racial discrimination).

Figure 13: 2016 news stories about racism in Google searches "unprofessional hair" and "three Black teenagers[6]."

A different example of social biases in search engines can be found in the autocomplete functionality. The aim of autocomplete is to help users in choosing keywords that are most likely to lead to the desired result, and in preventing users' spelling mistakes. The autocomplete suggestions can reveal the most frequent questions and the most common searches related to a specific topic.

Google search autocomplete suggestions often perpetuate negative stereotypes for queries related to gender, race, religion. A Google search in Turkish conveyed stereotypes in autocomplete about Greeks (Figure 14b) such as 'neden Yunanlılar türkleri sevmez/Why Greeks don't like Turks' (reflecting unfairness - racial discrimination). Likewise, a google search in Greek conveyed stereotypes about Turks (Figure 14a) such as 'γιατί οι Τούρκοι μισούν τους Ελληνες/Why do Turks hate Greeks' (reflecting unfairness - racial discrimination). Autocomplete is a black box and there is no mechanism to explain the suggestions.

Figure 14: Stereotypes about Greeks (left) and Turks (right) as conveyed by Google Auto-complete to users based in Cyprus.

[6] https://www.theguardian.com/commentisfree/2016/jun/10/three-black-teenagers-google-racist-tweet

Social Media News Feeds

In the case of social media news feed, an example of bias is the chatbot Tay. Microsoft created Tay, a "teenage girl" AI chatbot. Tay became "an evil Hitler-loving, incestual sex-promoting" bot, which resulted in Microsoft deleting it (Figure 15). Tay said among other things that "Hitler did nothing wrong" and called their followers "daddy" (reflecting unfairness - gender / racial discrimination). Microsoft commented that "The AI chatbot Tay is a machine learning project, designed for human engagement'—the more she talks with humans the more she will learn".[7] It can be noted that there was no mechanism provided for explaining/interpreting the system's actions to users.

Figure 15: 2016 news stories about Microsoft's Tay chatbot.[8]

Another case of existing biases in social media News Feed is that of political microtargeting. Political microtargeting predicts whether a user is more likely to be receptive to an advertising, using users' information, online behavior and psychological characteristics. The essential difference between basic targeting and microtargeting is the method of broadcasting. In microtargeting something is broadcast to only those who actually are interested in it, in contrast with basic targeting which broadcasts something to everyone (Figure 16). Political microtargeting and the use of an individual's information for advertising poses a threat to electoral democracy. An example of political microtargeting is depicted in Figure 17. In the United Kingdom, data including opinions on topical issues, cookies and other available data used to determine whether or not to send voter campaign materials and, if so, to tailor the messages contained within it (reflecting unfairness - beliefs discrimination). Due to lack of public scrutiny, the consequences of political microtargeting could be politicians being able to promise everyone what they actually want, but without any intention to do it.

[7] https://www.vice.com/en_us/article/kb7zdw/microsoft-suspends-ai-chatbot-after-it-veers-into-white-supremacy-tay-and-you

[8] https://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/

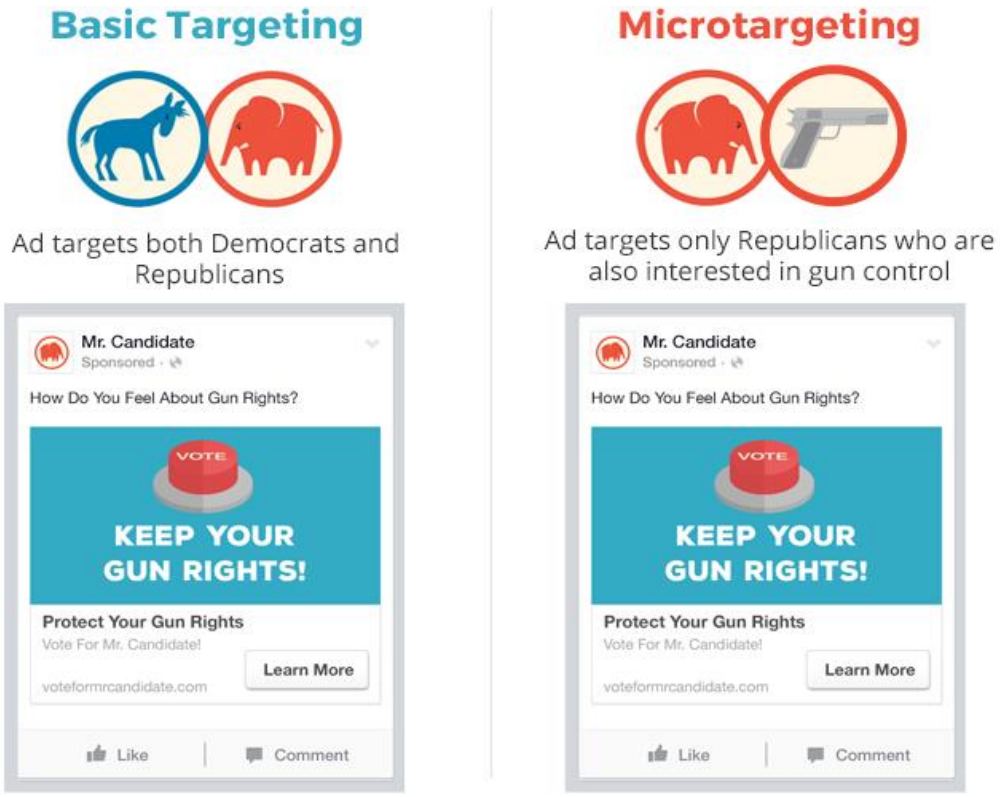


Figure 16: Facebook's microtargeting and its comparison to basic advertising.[9]

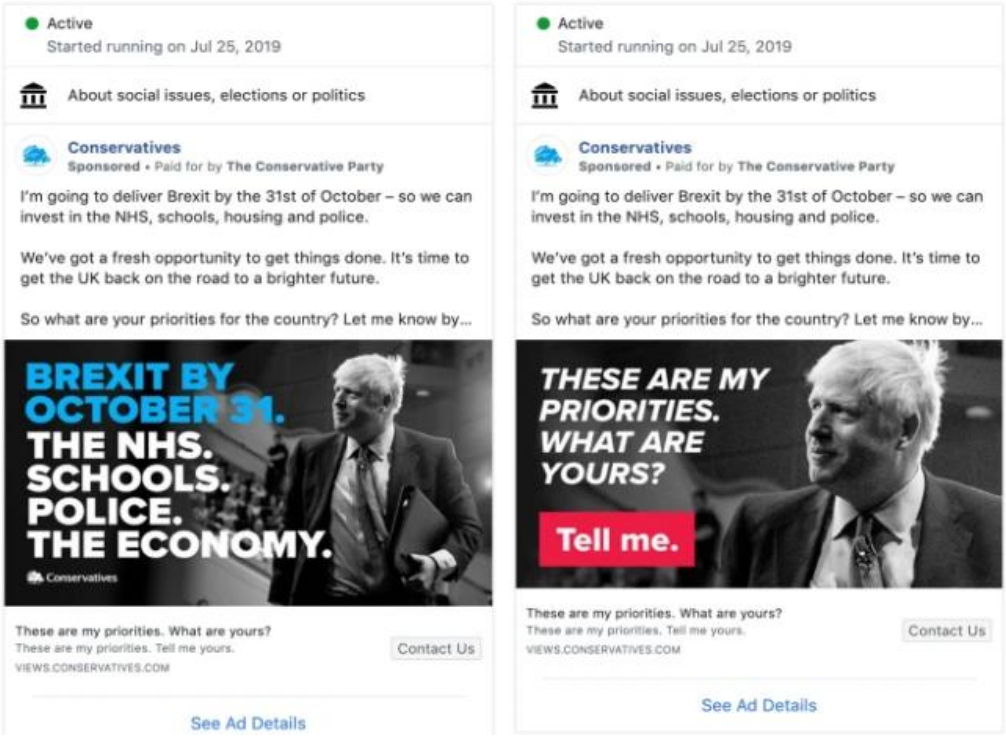


Figure 17: Ad for Brexit with tailored messages [10]

[9] http://fellows.rfiea.fr/dossier/comment-les-big-data-redessinent-l-avenir-de-la-democratie-et-de-l-etat-providence/article?language=en

[10] https://techcrunch.com/2019/08/05/uk-watchdog-eyeing-pm-boris-johnsons-facebook-ads-data-grab/

A different example of algorithmic bias in social media can be found within the LinkedIn search function. A LinkedIn search for a female contact results in the system asking the user if she was looking for a male with a similar name. Figure 18 depicts a search for the name 'Stephanie Williams'. In this example, LinkedIn presents a message to the user asking if she meant to search for 'Stephen Williams,' despite the fact that there are about 2,500 profiles with the name 'Stephanie Williams' (reflecting unfairness - gender discrimination). LinkedIn's spokeswoman said that the platform's suggestions are generated based on previous searches and how people are using the platform. LinkedIn has no mechanisms for explaining/interpreting the female-to-male prompts to its users.

1 of 2 | Searching 'Stephanie' on LinkedIn Searches for some common female names on professional social networking site LinkedIn bring up a prompt asking if users meant to look for similar-looking male names. ***Less***

Figure 18: LinkedIn's search function was found to exhibit gender bias.[11]

3.4.2 Sources of Bias

Having seen several examples of social and cultural biases in information access systems, we now classify algorithmic biases on the basis of their cause / source: i) data bias, ii) processing bias, iii) human bias.

Data Bias

Data biases can appear either in the training data (D) that is used to create an algorithmic model in an information system, as well as in the input (I) that a user submits to the system. It is natural that the social biases and discriminatory attitudes people hold in their offline lives would make their way into training data as well as system input. In other words, sensitive information about people (e.g., based on their gender, religion, age, sexual orientation and other sensitive attributes) may be expressed in the data. This can result in the system learning some unfair and discriminatory behaviors.

For instance, a user who holds traditional beliefs about gender roles, would be likely to perceive images of women nurses more relevant to a search for images of a "nurse." Given that search engines collect a massive amount of data from user interactions, gender beliefs that are prevalent in society will be reflected in the data collected, which in turn is used to train the algorithms behind the search engine. Likewise, users' social biases are reflected in the topics on which they search as well as the manner in which they formulate their search queries (e.g., "male nurse" implies that "nurse" alone will not return images of men and that men are typically not nurses). Finally, their biases and beliefs also influence the data they share on the Web and social media.

[11] https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias/

The case of the Microsoft Tay chatbot also illustrates the influence of data biases. In this case, the chatbot on Twitter interacted with the public, learning from the data (i.e., textual tweets) that users posted on the account. As shown in Figure 15, the chatbot - through its profile picture and tone - was depicted as an attractive teenage girl. This arguably was a catalyst for members of the public to post sexually explicit and sexist material, which was then used as training data for the bot.

Processing Bias

The second source of algorithmic biases is the manner in which algorithmic models are developed. Processing biases are those that appear while the algorithmic model is being learned / created / updated. Some algorithms are designed to *explicitly* use the sensitive attributes of persons for their predictive power. For instance, an algorithmic model for serving up online advertisements (a type of recommendation algorithm) for criminal defense legal services, might use information concerning an individual's race in its process. However, a more likely case is that the algorithm might use other less sensitive attributes (e.g., geographical location of the individual, or where he or she shops for groceries) to make inferences about whether or not the individual would be interested in the advertisement. The problem is that such attributes can be related to sensitive attributes, resulting in implicit discrimination. For instance, in many places, where a person spends most of her time is related to her social group (race or ethnicity, socio-economic class).

Human Bias

Finally, algorithmic biases may also be the result of humans, through inappropriate system development or usage. Below, we discuss the possibility for three types of stakeholders to cause human biases.

- Third party bias: Third parties, such as regulators or even other system users, can cause biases. For instance, a regulator in a non-democratic country, might constrain a system (e.g., a search engine) to repress outputs that are not in line with local law and/or cultural norms. A concrete example is that in France, it is illegal to buy or sell Nazi memorabilita; related system outputs are often suppressed. Other users may argue that this is a form of censorship and cultural bias. Another example can be drawn from recommendation systems, in which users rate content, which is then used to inform the algorithmic model, in an effort to improve all users' recommendations. However, users who hold explicit biases (e.g., are racist and intentionally provide low ratings to content related to persons of a minority race), have the potential to introduce their own human bias into the system.
- Developer bias: Those who develop algorithmic systems make many choices during the process, from the selection of the training data to the choice of the learning algorithm to be applied to the data, to the manner in which the algorithmic model is evaluated for its performance. Developers may inadvertently mishandle data and/or fail to see the point of view of the system's end users. Likewise, their own worldviews and biases (e.g., social stereotypes that influence their judgement) may affect their choices.
- User bias: Finally, users do not always use algorithmic systems appropriately. One common issue is "transfer context bias." This is when a user applies an algorithmic system in a context that is different from its intended use. An example could be a system for evaluating bank loan applications, which was designed and developed in the American context. Applying this system uncritically in the Cyprus socio-economic context could result in unexpected problems.

Finally, users should be aware that their own behaviors can affect those of an algorithmic system. Within an information access system such as a search engine or recommendation system, users are often unaware that their own behaviors are being tracked and used to provide feedback to the system. What this means is that if users uncritically accept results that are inappropriate or irrelevant (i.e., by uncritically choosing each time the top-ranking results presented to her), the system will not learn to improve. Popularity biases often develop in this manner. Top-ranking results tend to remain popular in search and recommendation systems, not because they are necessarily the best outputs, but because users do not take the effort to look further down the list of results in order to discover new or alternative system suggestions / outputs.

3.5 Addressing Bias in Algorithmic Systems:
Promoting Fairness, Accountability and Transparency (FAT)

In this section, we discuss how different sets of stakeholders can play a role in promoting algorithmic systems that treat people *fairly*, and that which can be held *accountable* for the decisions that they take, because their behaviors are *transparent* (i.e., can be interpreted and/or explained to a person). As depicted in Figure 19, at least three classes of stakeholders are involved in this process, and their activities are all necessary to ensure that systems are truly fair: Regulators, Developers, and Users.

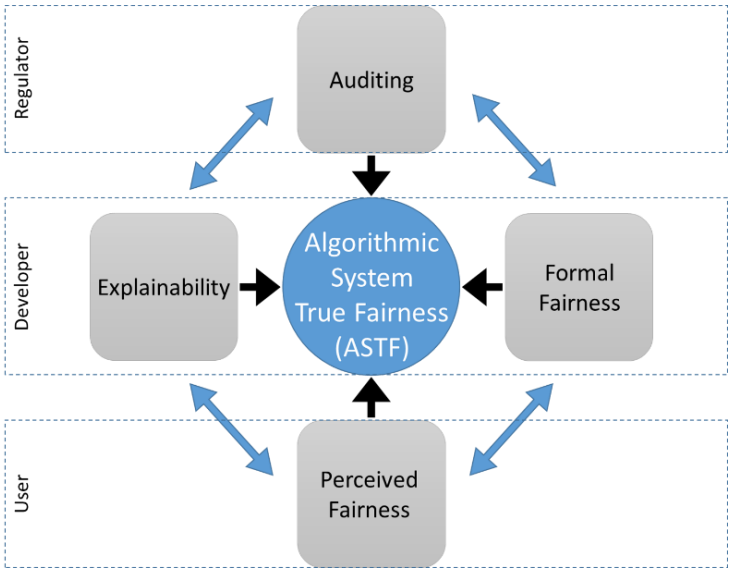


Figure 19: Processes and stakeholder roles in promoting FAT algorithmic systems.

Regulators (and other Auditors)

By *Regulators*, we refer to a neutral third party that is not involved in building the algorithmic system, but rather, is charged with evaluating its behaviors. Here, we refer to *Auditing* which, according to the Merriam-Webster dictionary is "a methodical examination and review,"[12] of an algorithmic system's behaviors, with

[12] https://www.merriam-webster.com/dictionary/audit

a focus on the detection of discriminatory behaviors. Researchers are developing methods for conducting various types of algorithmic audits, depending on the characteristics of the system in question.[13]

"Regulators" are typically associated with governments, such that they can actually take an action to hold system owners *accountable*. An example in the European Union is the enforcement of the General Data Protection Regulation (GDPR). According to Articles 51-59, Members States must have an independent and public authority for monitoring compliance and addressing non-compliance (i.e., the Information Commissioner's Office).[14] Several articles with GDPR (e.g., Article 22) place limits on systems that take automated decisions based on citizens' personal data.[15] Algorithmic IA systems, which use algorithmic user profiling, appear to be subject to such regulation. However, it remains to be seen how such regulations will be enforced.

In the near future, it is expected that industry bodies will also serve to audit and regulate algorithmic systems. For instance, the IEEE Standards Association currently has a standard under development for "Algorithmic Bias Considerations."[16] One in place, organizations will be able to become certified, demonstrating that they have done due diligence in terms of minimizing algorithmic bias in the systems they develop.

It should also be noted that we often observe other parties, such as journals and researchers, conducting audits. In this case, the goal is not enforce legal regulations but rather, to raise awareness of discriminatory behaviors in algorithmic systems. For example, racial bias in the COMPAS system, used in the United States to aid judges in determining criminal sentences, was first uncovered by journalists,[17] leading to several legal actions against its use.

Developers

Those who develop algorithmic processes and systems are usually the only stakeholders for whom the access to the code is a given. Even so, in many cases, the complex nature of the processes entails a need to conduct *Discrimination Detection*. To this end, machine learning researchers and practitioners have developed various technical procedures and tests to evaluate the fairness of their algorithms. These tests are *formal fairness* evaluations, in the sense that they are defined procedures - often being formalized as algorithms in and of themselves - to evaluate the extent to which the system's behavior discriminates against certain individuals or social groups. If an algorithmic system is found to be free of discrimination, it is said to have passed an *internal fairness certification*.

Another issue being addressed by developers is that of *explainability*. If an algorithm is developed such that its behaviors cannot be explained or interpreted by a human, there is little one can do to ensure its fairness. Therefore, in this sense, explainability is a necessary means to the desired end (i.e., fairness). As depicted

[13]Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, *22*.

[14] https://www.wired.co.uk/article/what-is-gdpr-uk-eu-legislation-compliance-summary-fines-2018

[15] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-does-the-gdpr-say-about-automated-decision-making-and-profiling/

[16] https://standards.ieee.org/project/7003.html

[17] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

in Figure 19, these development-focused processes are interrelated to ensuring the true fairness of an algorithmic system.

Users

The third stakeholder is the *User* of an algorithmic system. It is important to realize that even if an algorithmic system has been cleared through a formal process / test as being fair, it is not always the case that the user agrees. Thus, there is a kind of *informal fairness*, which concerns the user's perception of the system and its behaviors towards people.

The user's perception of fairness is, of course, important as it correlates to how much she trusts or distrusts the system and its output. The users are diverse, and they have differing perceptions of fairness that are in some cases shaped by their own beliefs, socio-cultural identities, life experiences etc. Researchers are currently trying to understand what kind of system explanations, as well as the kinds of fairness certifications, can help foster an appropriate level of trust of the system, in the user. Figure 19 depicts this interrelationship.

Educators

Educators such as yourselves, constitute a specific user group, as your own experiences with algorithmic IA systems will indirectly impact those of others. As will be described in this document, you play a key role in raising your students' awareness of the social and cultural biases that often surface in popular IA systems, and even those that you may use in the classroom. This *Guide* aims to help you integrate state-of-the-art findings from the research on fairness in algorithmic systems, into practical activities that you can use in fostering media-related algorithmic literacy amongst your students.

4. Rationale

The *New Oxford American Dictionary* defines 'literacy' as "the ability to read and write", and also the "competence or knowledge in a specified area". The European Commission, UNESCO, and other organizations emphasize that literacy today is also about one's ability to use ICTs and social media. To understand 21st century literacies, however, we need to acknowledge that algorithms and Artificial Intelligence (A.I) have increasingly big implications for our lives, especially our freedom, privacy, and access to opportunity. Hence, it's important to develop *algorithmic literacy*, namely, to raise awareness about the role that algorithms play and to push for a public accounting of their impact on our lives.

The rationale behind this *Guide*, then, is to raise teachers' awareness of the need to develop students' critical skills in understanding how algorithmic processes in AI systems such as Google Search, shape their view of the information landscape. The *Guide* includes activities (lesson plans) that raise awareness of the social and political consequences of algorithmic biases in the Cyprus context, which can be used in your own classrooms. A growing number of research studies show that although there is a perception that algorithms and artificial intelligence are objective and neutral, most of these algorithms are not only unknown to the public but also many of them have been shown to be *biased*.

Where does this bias come from?

It's simple. Those who prepare datasets and/or develop these algorithms are humans, who may have their biases and be unaware of it. These biases may be against people of color or other minorities, women, or

individuals with special needs. Therefore, it is important for teachers and students to be aware of the different AI systems used in the classroom that rely on algorithms; understanding how these algorithms work and paying attention to how different groups of students are experiencing them is important to make informed decisions.

The general pedagogical principles of this effort are the following:

1. The teacher has the fundamental task of guiding students and facilitating learning through good pedagogical strategies e.g., explaining concepts, giving examples, providing space for argumentation, highlighting points, asking questions, giving feedback, asking students to perform specific tasks.

2. The teacher uses specific criteria to evaluate the pedagogical activities he/she develops and implements such as:
   - To what extent is the activity embedded in a real-world context?
   - To what extent do the students have opportunities to observe and reflect upon the tasks they are engaged?
   - To what extent is the cultivation of algorithmic literary contributing to the attainment of specific goals (see point #3)?
   - Do students become more skilled in identifying bias with the use of algorithms and proposing solutions?
   - To what extent do students spend time engaged with other students on a task?
   - To what extent do students improve in their ability to negotiate solutions for algorithmic transparency with other students?
   - To what extent are students engaged with the ethical, social and political implications of algorithmic bias?
   - Do the activities in which students are involved foster the generation of multiple complex solutions on algorithmic bias that can be analyzed and evaluated for their effectiveness?

3. The teacher cultivates civic algorithmic literacy, that is, values and practices that will prepare students for using social media and the Internet to address algorithmic bias and its ethical, social and political implications. The role of the teacher is to create pedagogical spaces for meaningful and critical engagement with algorithmic bias in the classroom, opening up avenues for specific action taking and working, at a pragmatic scale, to address algorithmic bias in ways that promote the public good.

5. Lesson Plans

5.1 Objectives

The following Lesson Plans are designed for Secondary Education level and aim at introducing students (ages 14-18) to basic algorithmic processes in order to raise their awareness of algorithmic bias. More specifically, the Lesson Plans will help students to:

1. Identify the widespread application of algorithms in their daily lives and understand how their everyday activities depend on algorithmic processes.
2. Demonstrate how algorithmic bias may affect their choices or decisions and employ strategies to explore how algorithmic manipulation works.
3. Elaborate critical arguments for the importance of algorithmic transparency and connect examples of algorithmic bias to issues of privacy, financial profit and social equality.

5.2 Activities and materials

5.2.1 Algorithms in our daily lives

Algorithms are behind many decisions we take in our everyday lives: we rely on algorithms to determine what to buy, where to eat, whom to be friends with and even whom to date! Therefore, understanding what an algorithm is and how algorithms work means that we have a better insight into how our everyday lives depend on algorithms.

The goals of this theme are to:

1. Provide basic definitions of algorithms.
2. Explore the widespread use of algorithms in our daily lives through specific examples of common digital instruments.
3. Explain how algorithms are used to collect data from users in order to refine their output.

Task 1: What is an Algorithm?

An algorithm is a list of rules to follow in order to solve a problem. Algorithms specify sequential steps that need to be taken in order to achieve a desired result.

Discuss with students the definition of algorithm using examples of algorithms we use in our everyday life (e.g., making a cake, thinking of a set of directions to get to the park, getting dressed in the morning, following a list of instructions to make a table).

However, the digital tools we use in our everyday lives are powered with more complex algorithms that allow us to be more efficient in our decision-making. These algorithms reduce the complexity of information around us into a few choices that appeal to us and our interests. For instance, when Amazon recommends books for us based on what we have bought it uses an algorithm to figure out what other books were read by those who made similar purchases. Another example is when Facebook recommends us friends. It looks into our friends' friends in order to suggest that we may want to be linked to those individuals, too.

Algorithms are used in all areas of computing. Give students examples of how algorithms and then ask them to think of other examples of such algorithms. Algorithms examples are:

- Google's search engine which uses an algorithm to find the best matches for search terms. This algorithm decides which pages are listed first when you search for something.
- The Amazon website uses algorithms to decide the find results and set the price of products.

Task 2: How/Why do algorithms create bubbles?

Algorithms aid our decision-making by helping us get what we want because we keep telling them who we are and what we like. Algorithms are what makes our devices "smart." They make it seem as if a machine (computer, phone, tv) is thinking like we are thinking. In reality, however, it is the algorithm that we have "trained" with our behavior, our choices, our picks that makes the machine able to know or predict what we would like to do. Our habits, our preferences, our "likes" are part of the process that makes the algorithms work and become more efficient.

However, if we continuously feed into an algorithm only our personal preferences/likes/choices/picks then what we get back is something that reinforces all of these personal preferences. This creates a bubble around us.

- Each person fills out their personal filter bubble, what they think might be inside (and therefore outside) their bubble. For example, person X thinks that the items she recently purchased are inside her bubble.

Task 3: What types of personal data are we willing to disclose in order to receive personalized results?

Given the ubiquity of algorithms in our lives, it is important to ask the question: What criteria do algorithms use to decide results that are more relevant to us?

- Use Data Cards for exploring how decisions are made by algorithms and the impact that they have on our lives. Use the card to start a conversation on the information that we provide as input to these algorithms. We often provide these data willingly but oftentimes information about us is revealed through our activities.
- Ask students to think of how "personalized" advertising works. Provide examples of personalized advertising that is both solicited (asking Amazon to recommend books for you) and unsolicited (an ad for a hotel in Brussels that appears on a news website after you have searched for airline tickets to Brussels).
- Discuss with students the different types of personal data (e.g., age, gender, religion, income, education, occupation), what types of data are worth more to us and to the companies that use the data. Ask students to think of the following questions:
    - What information have you shared when using some of these search engines/digital applications?
    - What kind of information do you consider more valuable?
    - Which information about you do you think is more valuable to companies?
    - What do you know about how companies collect, use this data?

Recommended Videos for Discussion:

- Shoshana Zuboff on surveillance capitalism | VPRO Documentary
- Free is a Lie
- Living with Algorithms; Why should you care about algorithms?

5.2.2 How Algorithms work

Algorithms "learn" and become more efficient; about the questions we ask them. When we ask certain questions and then show preference for certain answers/results/outcomes, then the algorithm "learns" that this is what it should keep giving us back. We create the reality that the algorithms deliver to us. Algorithms reflect the world we are showing them. When we use algorithms to view, understand and represent people, then the algorithm will automatically turn people into categories, profiles and types.

In other words, when the data that are used to "train" the algorithm is biased, or it represents a particular limited view of a society or a community or the world then the result is also biased. This result is then used to make further decisions. Therefore, a feedback loop is created that is difficult to break. The question has arisen is how does this feedback loop affect our lives and our future if we are not aware of it?

The goals of this theme are:

1. To investigate the different expressions of algorithmic bias.
2. To explore how the input of different variables alters algorithmic results.
3. Demonstrate how the use of algorithms can perpetuate social problems such as prejudice, intolerance and discrimination.

Task 4: Demo of the Filter Bubble

Algorithmic systems use personal data to automatically filter and/or rank content based on the user profile, to guide users to the most relevant material.

- The search engine uses variables such as your location and your language to provide results. If you are signed into the search engine, then other information such as your gender and age are used as a way to determine best results for you.
- Also, the search engine uses other data that you have unwittingly provided (your geographic location and movement, your previous searches, your preferences for certain websites) in order to determine what you might like to see at the top of your search results.

The education demo “Filter Bubble” was developed to demonstrate how the algorithms outputs could differ considering the input. The demo has an “Explicit" and an “Implicit" mode. The “Explicit" mode uses basic information given actively by the user, such as gender and age. The “Implicit" mode uses information inferred from the wording of the query, such as the first language of the user (implying their location, cultural background, etc.) or the adjectives/adverbs they’ve used (implying their goals or attitudes). Students can use the Filter Bubble and explore how the variation of their input affects the results provided by the search engine.

Students can select a “search query” from a list in the first page of the demo and interact with the results of their search. For instance, if the topic is Cyprus and the original query is the name in English (“Cyprus") while the variations are in different languages (Greek, Turkish, and Russian). Clicking a variation modifies the user model used to filter the results, updating the text or images shown. After the student has explored at least one variation, the button to move onto the Explanation page is enabled. On the Explanation page students can find explanations for the filter bubble effect, including: a short video that demonstrates the filtering and reordering process with images, an interactive section imitating the search results (allowing the user to change the characteristics of the user profile and see the changes immediately), and text with dynamic phrases that depend on the user’s profile and topic. There are also some tips for understanding and managing personal information given to a platform.

Task 5: Discussion

Search engines and social networks filter out information and data about us that present us in a particular manner, depending on who they think we are.

Is this positive or negative?

✔ We are exposed to information and choices found close to our previous interests (as those have been registered before in our online history).

✘ We are enclosed in a personal sphere of information, where the same persons, information, news, and interests are recycled.

✗ Companies and governments have at their disposal detailed information about us => we are under constant surveillance.

Consequences of living in a Filter Bubble:

- We are not exposed to information that the search engine/algorithm calculates as an outlier or as 'awkward' information.
- We do not accept new, different stimuli and information (bias).

Encourage students to think of how they would create their own Filter Bubble in order to demonstrate algorithmic bias.

5.2.3 Algorithmic Transparency

Algorithms are a set of steps leading to an outcome. For instance, an algorithm people use in everyday life is a recipe. Algorithmic transparency is having access to all steps of the algorithm. Think of a recipe; an example of a fully transparent algorithm; the cook is given access to all of the necessary ingredients and steps that will lead to the desired tasty food.

The goals of this theme are:

1. To explore the question of algorithmic transparency and what it might mean for everyday users.
2. To critically debate ethical dilemmas in the use of algorithms (privacy, pursuit of profit, discrimination)
3. To develop standards and guidelines for the use of digital tools that work with algorithms.

Task 6: Participants' views on algorithm transparency.

Ask students their views on algorithmic transparency.

- Is algorithmic transparency important? If so, why?
- What might algorithmic transparency look like? How should this be communicated to everyday users of digital tools?
- Discuss recommendations that participants might have in relation to changes that could be made to the way that the Internet currently functions.

Possible task with the following sources from UNBIAS website (e.g.,What can parents do? Basic ideas about the GDPR; principles for accountable algorithms).

Further Sources for Algorithmic Transparency Debate:

- Principles for Accountable Algorithms and a Social Impact Statement for Algorithms
- Child Safety Online: A Practical Guide for Providers of Social Media and Interactive Services
- Guide to the GDPR

Further sources for internet and discrimination

- Google under fire over 'racist' image search results for 'unprofessional hair'
- "I think my blackness is interfering": does facial recognition show racial bias?
- HP Investigates Claims of 'Racist' Computers

Practical tools for the reflective use of the social media

- incognito browsers
- use of many search engines

- use of different services (not the same company, not necessarily commercial services)
- reflection when exposed to automated "suggestions" or ads: we ask: "why am I seeing this ad?"

As digital citizens, we demand our rights

- we demand transparency; we are entitled to know what they know about us
- we demand to know who decides and how decisions are made concerning the ways data about us is going to be used
- we demand to have a saying in these decisions

5.3 Evaluation

The goals of the Evaluation are to:

1. Monitor students' learning (establish the development, strengths and weaknesses of each student)
2. Provide feedback to educators to improve their teaching
3. Provide feedback to students to improve their learning.

Educators can evaluate students' performance using both continuous (Formative) and Summative assessment. During the execution of the Lesson Plans educators can carry out activities—either individual or group activities—in order to assess the students' level of comprehension and the need for further clarifications.

Formative assessment is more diagnostic than evaluative. More specifically, the goal of formative assessment is to help students identify their strengths and weaknesses and target areas that need work. Also formative assessment aims to help educators improve and adjust their teaching methods by recognising where their students are struggling.

On the other hand, the goal of summative assessment is to evaluate students' learning and their academic achievements at a specific time (e.g. end of the lesson, end of the academic year) by comparing it against some standards or benchmark. Furthermore, Summative assessment aims to identify common gaps in students learning and recognise the weaknesses and the strengths of the lesson plans. Finally, it helps educators to identify if there is a need to develop further activities and change teaching methods.

Examples for Formative assessment include:

- Ask students to create a visual map of what they learnt
- Small quizzes after each class
- In-class discussion at the end of each class
- Homework assignments with structured feedback

Examples for Summative assessment includes:

- A "final" exam at the end of the Lesson Plans
- A group project. Create groups of 4-5 students and ask each group to present a specific topic in class (e.g. they can choose an IA system they commonly use and try to identify the information it collects and discuss if they perceive the system as fair). At the end of each lesson give time to students to work on their project considering what they learnt in the class.

6. Conclusion

In this deliverable, we presented an easy-to-use guide for secondary school teachers working in Cyprus. The guide has been developed to help teachers to raise their students' awareness of algorithmic processes and algorithmic bias. The guide focused on explaining the problem of algorithmic biases and provided an overview of the techniques that researchers have proposed to mitigate the bias. The document summarized the core technical concepts surrounding algorithmic system transparency. In particular, the framework which was presented in D4.1 has been presented in a manner that is understandable and useful for the secondary school teachers. Furthermore, examples of biases in biases in Algorithmic IA Systems and their causes have been presented followed by ways to promote algorithmic systems that treat people fairly. Finally, this document has provided lesson plans to the teachers presenting the objectives, materials, activities, and finally, the evaluation which the teachers can use to develop their own lessons for their classrooms. The document will be used for the development of the teacher training sessions, which will be carried out in WP5.

Annex - Greek Translation

1. Εισαγωγή: Ο σκοπός του οδηγού

Ο οδηγός αυτός στοχεύει να σας βοηθήσει να προωθήσετε την αλγοριθμική παιδεία (algorithmic literacy) στην τάξη σας. Οι αλγοριθμικές διαδικασίες που είναι τεχνικά πολύπλοκες και διαδραματίζουν πολύ βασικό ρόλο στις τεχνολογίες πρόσβασης που χρησιμοποιούμε καθημερινά. Είτε ψάχνουμε για πληροφορίες στο διαδίκτυο, είτε ψάχνουμε στο NetFlix μία ταινία να παρακολουθήσουμε, είτε κοινωνικοποιούμαστε σε μία πλατφόρμα κοινωνικών μέσων, αλγοριθμικές διαδικασίες διαμεσολαβούν σχεδόν σε όλη μας την πρόσβαση σε πληροφορίες. Έρευνες έχουν δείξει ότι πολλοί χρήστες δεν αντιλαμβάνονται την παρουσία αυτών των αλγορίθμων και πιστεύουν ότι έχουν πρόσβαση σε όλες τις πληροφορίες που υπάρχουν σε μια συγκεκριμένη πλατφόρμα/περιεχόμενο. Επίσης, συχνά παραβλέπετε το γεγονός ότι οι τεχνολογίες δεν είναι ουδέτερες. Ενσωματώνουν και αντανακλούν τις ανθρώπινες αξίες (π.χ. σε μία μηχανή αναζήτησης, την αποδοτικότητα και την ταχύτητα, αλλά και την προτίμηση για συγκεκριμένες πηγές πληροφοριών) και μπορούν με αυτό τον τρόπο να διαμορφώσουν τις απόψεις και τις πρακτικές μας. Για παράδειγμα, μία αναζήτηση στο Google για τη λέξη ‘nurse (νοσηλευτής/τρια)’ αντικατοπτρίζει τις ανθρώπινες πεποιθήσεις ότι υπάρχουν περισσότερες γυναίκες νοσηλεύτριες.

Για όλους αυτούς τους λόγους, είναι σημαντικό να γνωρίζουμε τον ρόλο που παίζουν οι αλγόριθμοι και πώς μπορούν να επηρεάσουν τις ικανότητες των πολιτών να παραμένουν ενημερωμένοι και αφοσιωμένοι. Ο οδηγός αυτός θα σας βοηθήσει να κατανοήσετε τη φύση των αλγοριθμικών συστημάτων πρόσβασης πληροφοριών και πώς μπορεί να προκύψουν κοινωνικές προκαταλήψεις μέσα σε αυτά τα συστήματα. Κατά τη διάρκεια της διαδικασίας, θα παρουσιάσουμε παραδείγματα συστημάτων τα οποία πιθανόν χρησιμοποιείτε τακτικά. Στην συνέχεια, θα επανεξετάσουμε τις παιδαγωγικές αρχές που καθοδηγούν την προσέγγιση μας για την προώθηση της αλγοριθμικής παιδείας. Τέλος, θα παρουσιάσουμε παραδείγματα αλληλεπιδραστικών δραστηριοτήτων στην τάξη που μπορούν να χρησιμοποιηθούν για την ευαισθητοποίηση των μαθητών για τον ρόλο των αλγορίθμων στην πρόσβαση στις πληροφορίες. Κατά τη διάρκεια των σεμιναρίων (πρόσωπο-με-πρόσωπο) που θα συνοδεύσουν αυτό τον οδηγό, θα σας ενθαρρύνουμε να προσαρμόσετε αυτές τις δραστηριότητες, διαμορφώνοντας ένα πλάνο μαθήματος που θα είναι αποτελεσματικό με βάση τις ανάγκες των μαθητών σας.

2. Ορισμοί και βασικές τεχνικές έννοιες:

2.1 Αλγοριθμικές διεργασίες και συστήματα:

Ένας αλγόριθμος, σύμφωνα με το λεξικό Merriam-Webster, ‘είναι μία διαδικασία βήμα-προς-βήμα για την επίλυση ενός προβλήματος ή την επίτευξη κάποιου σκοπού’. Είναι ένας τρόπος να οργανωθούν οι σκέψεις και οι ενέργειες που χρειάζεται να εκτελεστούν κατά τη διάρκεια μιας εργασίας για να οδηγήσουν σε ένα επιθυμητό αποτέλεσμα. Ένα απλό παράδειγμα είναι μία συνταγή, η οποία εξηγεί στον μάγειρα όχι μόνο τη λίστα με τα συστατικά και τα απαραίτητα σκεύη που χρειάζονται για να φτιάξει το επιθυμητό φαγητό, αλλά και τα βήματα που πρέπει να ακολουθηθούν και την κατάλληλη σειρά. Έτσι, μία συνταγή είναι ένας εντελώς διαφανής αλγόριθμος, που περιγράφει επαρκώς τις εισόδους (τα συστατικά, τα σκεύη), την επεξεργασία (τα βήματα που πρέπει να ακολουθηθούν, καθώς και οποιεσδήποτε υποθέσεις ή πράγματα που πρέπει ο μάγειρας να προσέξει) και το αναμενόμενο αποτέλεσμα (το πιάτο που παρασκευάζεται, ο αριθμός των μερίδων κ.τ.λ.)

Άλλα παραδείγματα αλγορίθμων στην καθημερινή ζωή είναι η ταξινόμηση ενός συνόλου αρχείων με βάση την ημερομηνία παραγωγής τους ή η συναρμολόγηση ενός αντικειμένου που αγοράσαμε (π.χ. ένα αντικείμενο από το ΙΚΕΑ, το οποίο αποτελείται από πολλά διαφορετικά μέρη και πωλείται με έναν ('πως-να') οδηγό συναρμολόγησης). Σε κάθε μία από αυτές τις περιπτώσεις, για να φτάσουμε στο επιθυμητό αποτέλεσμα (μία στοίβα αρχείων που είναι σε σειρά κατά ημερομηνία, ένα κομμάτι επίπλου που είναι συναρμολογημένο και πλήρως λειτουργικό), κάποιος πρέπει να γνωρίζει τις εισόδους στη διαδικασία, καθώς και τη διαδικασία βήμα-προς-βήμα που θα μας οδηγήσει στο αποτέλεσμα.

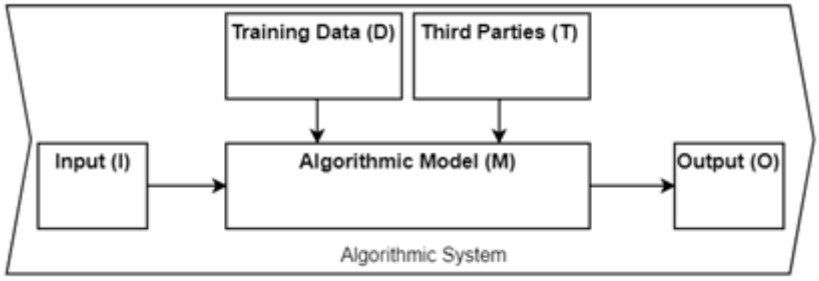
Τα *πληροφοριακά συστήματα (information systems)* βασίζονται σε ψηφιακά εφαρμοσμένους αλγόριθμους κωδικοποιημένους σε μία γλώσσα προγραμματισμού για να πραγματοποιήσουν λειτουργίες που οδηγούν σε μία επιθυμητή λειτουργικότητα ή / και τελικό αποτέλεσμα για τους χρήστες τους. Για παράδειγμα, σε ένα σύστημα μισθοδοσίας που υπολογίζει τους μισθούς των εργαζομένων μίας εταιρείας, υπάρχει ένας αλγόριθμος που προσδιορίζει όλες τις οδηγίες που απαιτούνται για την εκτέλεση αυτής της πολύπλοκης εργασίας. Αρχικά, το σύστημα απαιτεί όλα τα κατάλληλα δεδομένα εισόδου για ένα συγκεκριμένο υπάλληλο (π.χ. το ποσοστό αμοιβής και υπερωριών, τον αριθμό των ωρών που εργάστηκε σε μια δεδομένη χρονική περίοδο, καθώς και τις απαραίτητες προσωπικές πληροφορίες που απαιτούνται, όπως η οικογενειακή κατάσταση ή/και ο φόρος εισοδήματος, κ.τ.λ.). Στη συνέχεια, ο αλγόριθμος αυτός θα πρέπει να καθορίζει όλα τα απαραίτητα βήματα για τον υπολογισμό του ακαθάριστου μισθού (π.χ. πολλαπλασιάζει τον ωριαίο μισθό με το χρόνο εργασίας, πολλαπλασιάζει τον ωριαίο μισθό υπερωριακής εργασίας με οποιαδήποτε υπερωριακή απασχόληση και τα προσθέτει), τις διάφορες αποκοπές που θα πρέπει να κάνει ο εργοδότης (π.χ. φόροι, κοινωνικές ασφαλίσεις, κ.τ.λ.) και τέλος ο καθαρός μισθός που πρέπει να καταβληθεί στον εργαζόμενο.

2.2 Αλγοριθμικά συστήματα που μαθαίνουν

Τα πιο πάνω παραδείγματα αποτελούν *στατικές (static)* αλγοριθμικές διαδικασίες. Σε μία διαδικασία όπως μία συνταγή ή ένα σύστημα μισθοδοσίας εργαζομένων, ο αλγόριθμος είναι διαφανής και δεν αλλάζει αυτόματα.[18] Αντίθετα, οι αλγόριθμοι πίσω από τα σύγχρονα διαδραστικά συστήματα που χρησιμοποιούμε σήμερα για την πρόσβαση σε πληροφορίες είτε στο διαδίκτυο είτε στα κοινωνικά δίκτυα τείνουν να είναι πολύ δυναμικοί. Δηλαδή, έχουν σχεδιαστεί με τέτοιο τρόπο που τους επιτρέπει να μαθαίνουν και να βελτιώνουν συνεχώς την απόδοσή τους, με βάση τα δεδομένα που αντικατοπτρίζουν τις συμπεριφορές των χρηστών και τις προτιμήσεις τους κατά τη χρήση του συστήματος.

Η Εικόνα 1 απεικονίζει τη βασική αρχιτεκτονική ενός αλγοριθμικού συστήματος που μαθαίνει από τα δεδομένα (δηλαδή βασίζεται σε τεχνικές *μάθησης μηχανών / machine learning* techniques). Όπως παρατηρείτε, υπάρχουν πέντε στοιχεία σε ένα τέτοιο σύστημα. Αυτά τα στοιχεία εξηγούνται παρακάτω, χρησιμοποιώντας το παράδειγμα μιας μηχανής αναζήτησης (π.χ.Google ή Microsoft Bing).

[18] Φυσικά, υπάρχουν περιπτώσεις υπό τις οποίες θα χρειαστούν αλλαγές στον αλγόριθμο. Προκειμένου να εκτελεστεί με επιτυχία σε μεγάλο υψόμετρο, μια συνταγή για ένα κέικ θα χρειαστεί αναπροσαρμογή. Σε ένα σύστημα μισθοδοσίας, οι αλλαγές στη φορολογική νομοθεσία απαιτούν την ενημέρωση του αλγορίθμου του συστήματος. Ωστόσο, υποθέτουμε ότι αυτές οι αλλαγές γίνονται με μη αυτόματο τρόπο, δεν στηρίζονται σε πληροφορίες βασισμένες σε δεδομένα και δεν συμβαίνουν αυτόματα.



Εικόνα 1: Βασική αρχιτεκτονική ενός αλγοριθμικού συστήματος

- **Είσοδος (Input) (I):** όταν ο χρήστης χρησιμοποιεί το σύστημα, εισάγει κάποια συγκεκριμένη τιμή/ες για να εκτελέσει μια δεδομένη λειτουργία του συστήματος. Σε μία αναζήτηση στο διαδίκτυο, ο χρήστης παρέχει ένα σύνολο λέξεων κλειδιών, οι οποίες εκφράζουν την ανάγκη του για πληροφορίες σε σχέση με ένα συγκεκριμένο θέμα.[19]
- **Έξοδος (Output) (O):** η έξοδος είναι οι πληροφορίες που παράγει το σύστημα, με βάση τις εισόδους του χρήστη. Στην περίπτωση αναζήτησης στο διαδίκτυο αυτό είναι το ταξινομημένο σύνολο ιστοσελίδων που οι αλγόριθμοι θεωρούν ότι είναι πιο πιθανό να είναι χρήσιμοι για τον χρήστη.
- **Αλγόριθμος (Algorithm) (M):** το αλγοριθμικό μοντέλο ενός συστήματος είναι ο πυρήνας του. Αυτό είναι το μέρος του συστήματος που αφού έχει εκπαιδευτεί από τα δεδομένα, χαρτογραφεί μια δεδομένη είσοδο σε μια δεδομένη έξοδο. Στην περίπτωση μίας σύγχρονης μηχανής αναζήτησης, το αλγοριθμικό μοντέλο δεν είναι στην πραγματικότητα ένας μόνο ο αλγόριθμος, αλλά ένα ολόκληρο σύνολο αλγοριθμικών διεργασιών. Ωστόσο, ίσως ο πιο πολυσυζητημένος αλγόριθμος που χρησιμοποιείται σε μία μηχανή αναζήτησης είναι ο *αλγόριθμος κατάταξης*, ο οποίος αποδίδει μία βαθμολογία σε κάθε ιστοσελίδα που ανακτάται ως αποτέλεσμα στις λέξεις κλειδιά του χρήστη, έτσι ώστε οι σελίδες να μπορούν στη συνέχεια να παρουσιαστούν στο χρήστη κατά σειρά.
- **Δεδομένα Εκπαίδευσης (Training Data) (D):** τα δεδομένα εκπαίδευσης χρησιμοποιούνται για την εκπαίδευση του αλγοριθμικού μοντέλου (M) όταν εφαρμόζονται τεχνικές εκμάθησης μηχανών. Όπως αναφέρθηκε στην περίπτωση συστημάτων πρόσβασης στις πληροφορίες, όπως οι μηχανές αναζήτησης, τα δεδομένα εκπαίδευσης αποτελούνται από παρατηρήσεις που αφορούν τη συμπεριφορά των χρηστών, τις προσωπικές ιδιότητες και τις προτιμήσεις πληροφόρησης. Ένα παράδειγμα για την περίπτωση αναζήτησης στο διαδίκτυο, είναι το σύνολο δεδομένων κατάρτισης (training dataset) με σκοπό την ανάπτυξη ενός αλγορίθμου κατάταξης που θα περιέχει πληροφορίες σχετικά με τους τύπους ιστοσελίδων που οι χρήστες θεωρούν ότι είναι σχετικές με ένα συγκεκριμένο θέμα.
- **Περιορισμοί τρίτων (Third Party Constraints) (T):** σε μερικές περιπτώσεις αλγοριθμικών συστημάτων, ένα τρίτο μέρος (δηλαδή όχι ο χρήστης του συστήματος, ούτε και ο προγραμματιστής του) επιβάλλει ορισμένους περιορισμούς στον τρόπο που λειτουργεί το σύστημα. Αυτοί μπορεί να είναι οι ιδιοκτήτες ή οι χειριστές του συστήματος, οι ρυθμιστικές αρχές και άλλοι που επηρεάζουν τη χρήση και τα αποτελέσματα του συστήματος. Ένα παράδειγμα στην περίπτωση αναζήτησης στο διαδίκτυο, είναι όταν οι χειριστές πρέπει να περιορίζουν τις συμπεριφορές του συστήματος έτσι

[19] Πρέπει να σημειωθεί ότι πιο έμπειροι χρήστες συχνά εισάγουν περισσότερα από απλές λέξεις-κλειδιά, χρησιμοποιώντας προηγμένες διαμορφώσεις για να καθορίσουν την ανάγκη για πληροφορίες από συγκεκριμένες πηγές, σε μια δεδομένη γλώσσα κ.λπ.

ώστε να συνάδουν με τους νόμους κάθε χώρας. Οι μηχανές αναζήτησης, παραδείγματος χάρη, πολύ συχνά περιορίζουν τα αποτελέσματα αναζήτησης (την έξοδο) που εμφανίζονται στους χρήστες που αναζητούν πληροφορίες χρησιμοποιώντας το όνομα ενός πραγματικού ατόμου, προκειμένου να συνάδουν με τους κανονισμούς της Ευρωπαϊκής Ένωσης για την προστασία των προσωπικών δεδομένων.[20]

2.3 Συστήματα αλγορίθμων για πρόσβαση σε πληροφορίες

Έχοντας εξετάσει τα χαρακτηριστικά ενός δυναμικού αλγοριθμικού συστήματος, θα δούμε τώρα μερικά παραδείγματα συστημάτων πρόσβασης πληροφοριών (information access systems) με τα οποία εσείς και οι μαθητές σας αλληλεπιδράτε τακτικά. Συγκεκριμένα, πιο κάτω εξετάζονται τρεις κατηγορίες συστημάτων πρόσβασης (ΙΑ). Αρχικά, περιγράφονται γενικά από την άποψη της λειτουργικότητάς τους και των συστατικών τους, χρησιμοποιώντας την ορολογία και τις συντομογραφίες που παρουσιάζονται στην Εικόνα 1. Εν συνεχεία, παρέχονται παραδείγματα για την κάθε κατηγορία συστημάτων.

2.3.1 Συστήματα Σύστασης (Recommender Systems)

Αυτά είναι αλγοριθμικά συστήματα που παρέχουν συγκεκριμένες προτάσεις στους χρήστες κατά τη διάρκεια της αλληλεπίδρασής τους με το σύστημα. Η στοχευμένη διαφήμιση (ενώ χρησιμοποιείτε το διαδίκτυο ή μία εφαρμογή για κινητά τηλέφωνα) είναι ένα πολύ καλό παράδειγμα ενός συστήματος συστάσεων. Οι προτάσεις (π.χ. για επιχειρήσεις που βρίσκονται γεωγραφικά κοντά σε ένα χρήστη) παρέχονται με βάση το τι ξέρει το σύστημα για τον χρήστη (τα δημογραφικά χαρακτηριστικά, τις συμπεριφορές και τις προτιμήσεις του) και άλλες σχετικές πληροφορίες (τη γεωγραφική θέση, την ημέρα της εβδομάδας, την ώρα της ημέρας, τον καιρό, κ.λπ.). Στην πραγματικότητα όμως, οι προτάσεις που παρέχονται ενδέχεται να είναι ή όχι οι καλύτερες επιλογές για τους χρήστες. Αυτές οι συστάσεις μπορούν να είναι και χρηματοδοτημένες από τρίτους που έχουν πληρώσει (π.χ. ένας διαφημιζόμενος που πλήρωσε για να προβάλλονται οι διαφημίσεις του σε στοχευμένες ομάδες ατόμων).

Παραδείγματα συστημάτων σύστασης:

- Διαδικτυακή στοχευμένη διαφήμιση (π.χ. μέσω του Google AdSense)
- Προτεινόμενα βίντεο στο YouTube (‘Επόμενο’ βίντεο)
- Προτεινόμενες ταινίες στο Netflix (‘Δημοφιλή στο Netflix’, ‘Παρακολουθήσατε πρόσφατα’)
- Προτάσεις για ηλεκτρονική αγορά (π.χ. ‘Πελάτες που είδαν αυτό το προϊόν είδαν επίσης και αυτό…’)

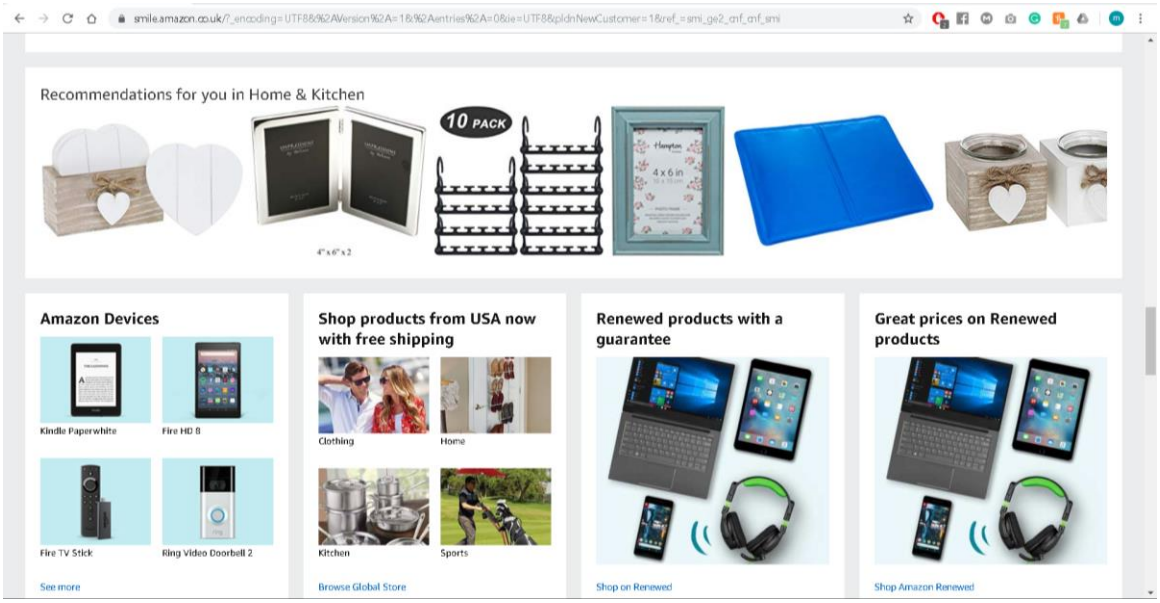
Σε ένα σύστημα σύστασης, η είσοδος (I) εξαρτάται από τη συμπεριφορά του χρήστη, με άλλα λόγια, αυτό που βλέπει ο χρήστης ή κάνει στο σύστημα τη δεδομένη στιγμή. Η έξοδος (O) του συστήματος είναι το σύνολο των προτάσεων που παρέχονται στο χρήστη. Τα δεδομένα εκπαίδευσης (D) περιέχουν τις προηγούμενες παρατηρήσεις των χρηστών, που θα επιτρέψουν στο αλγοριθμικό μοντέλο (M) να χαρτογραφήσει τα τρέχοντα χαρακτηριστικά του χρήστη (δημογραφικά στοιχεία, συμπεριφορές, τοποθεσία κ.ά.) σε αντικείμενα που ίσως τον ενδιαφέρουν. Τέλος, οι περιορισμοί από τρίτους (T) μπορεί να περιλαμβάνουν όχι μόνο αυτά που αναφέρθηκαν πιο πάνω αλλά και πρόσθετες πληροφορίες που παρέχονται από άλλους χρήστες (π.χ. βαθμολογίες σε ένα συγκεκριμένο αντικείμενο, οι οποίες θα

[20] Το “δικαίωμα να ξεχαστείς”.

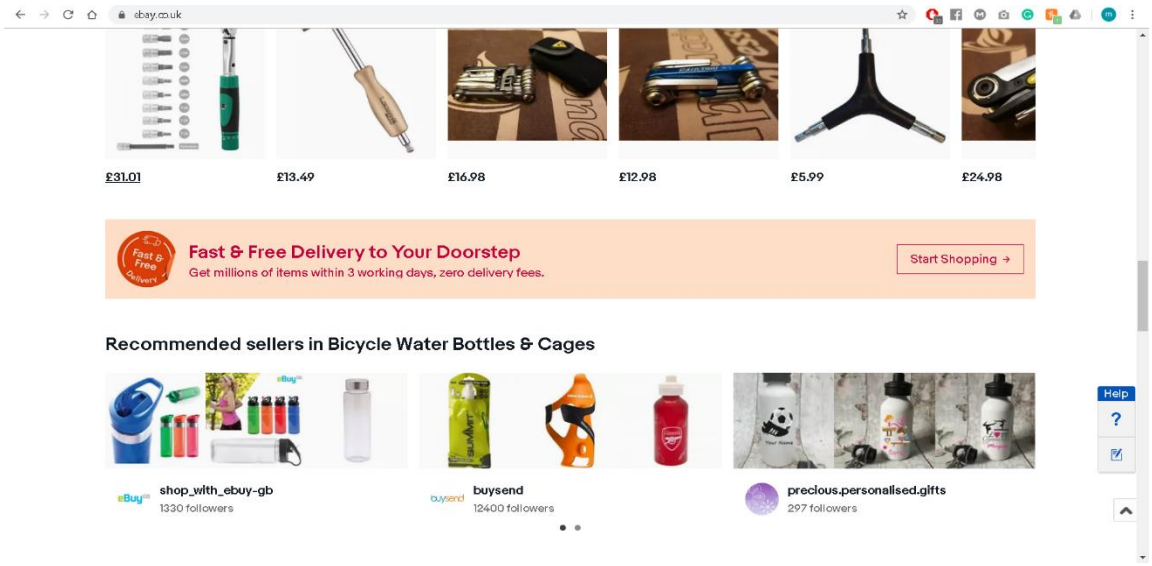
μπορούσαν να επηρεάσουν την αξιολόγηση του μοντέλου σχετικά με την ποιότητα ή την καταλληλόλητα του αντικειμένου).

Ένα παράδειγμα συστήματος σύστασης είναι το Amazon. Το Amazon εμφανίζει προτάσεις (O) στο χρήστη με βάση το προφίλ του και το ιστορικό παραγγελιών του (D). Η Εικόνα 2 απεικονίζει την αρχική σελίδα, ενός χρήστη στην Amazon. Ο χρήστης αγόρασε πρόσφατα μία κορνίζα από το Amazon.co.uk (I). Στην Εικόνα 2 φαίνονται οι σχετικές προτάσεις (O) από την Amazon στον χρήστη, που περιλαμβάνουν κορνίζες και άλλα σχετικά προϊόντα (διακόσμηση εσωτερικού χώρου).

Ακόμα ένα παράδειγμα ενός συστήματος σύστασης είναι το eBay. Η Εικόνα 3 δείχνει τις προτάσεις (O) που ο αλγόριθμος (M) του eBay δίνει σε ένα χρήστη που πρόσφατα έψαχνε μία φιάλη νερού (I).



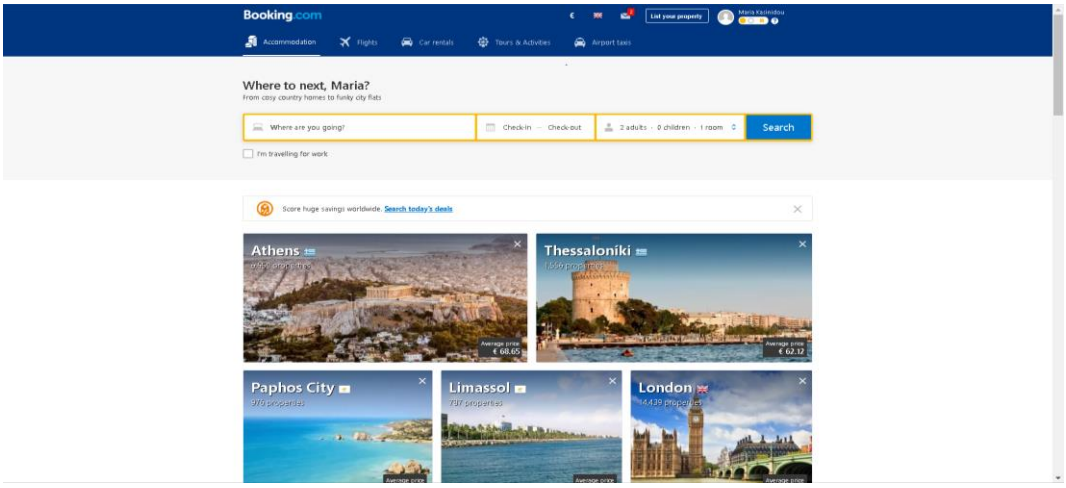
Εικόνα 2: Οι προτάσεις (O) της Amazon με βάση το προφίλ του χρήστη και προηγούμενες αγορές του (I).



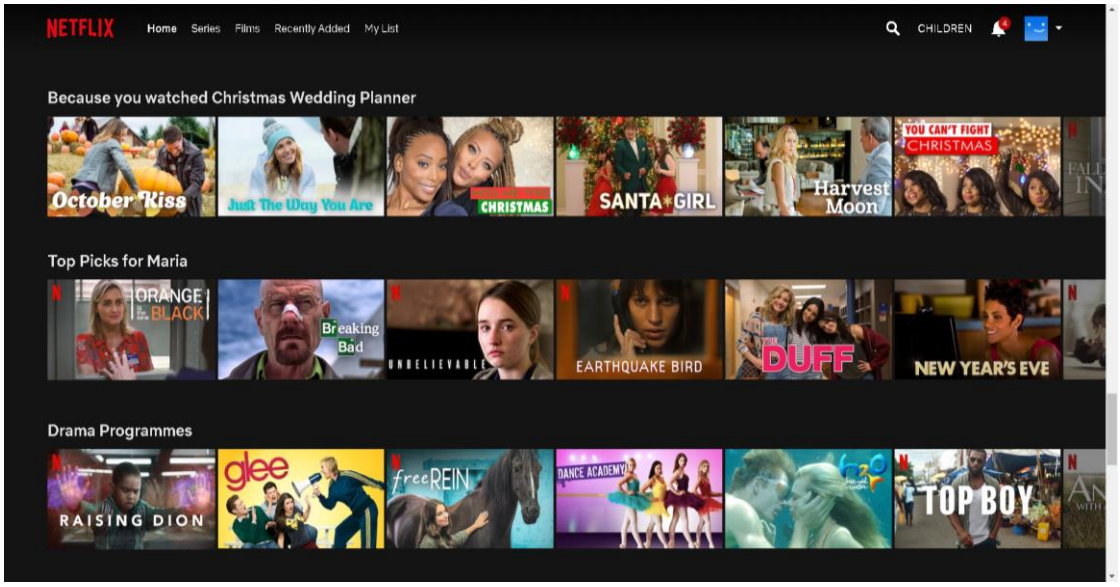
Εικόνα 3: Οι προτάσεις (O) του Ebay.co.uk με βάση τις αναζητήσεις του χρήστη και το προφίλ του (I).

Ένα διαφορετικό σύστημα συστάσεων είναι το Booking.com. Οι προτάσεις της ιστοσελίδας δεν βασίζονται μόνο στο προφίλ του χρήστη και στις πρόσφατες ενέργειες του (I), αλλά και σε άλλες πληροφορίες, όπως η γεωγραφική θέση του χρήστη (I). Ένας χρήστης που αυτή τη στιγμή βρίσκεται στην Κύπρο (I) λαμβάνει προτάσεις για κοντινές πόλεις όπως η Λεμεσός και η Πάφος (O) (Εικόνα 4).

Το Netflix είναι ένα ακόμα σύστημα συστάσεων που χρησιμοποιείται ευρέως. Οι προτάσεις ταινιών στο Netflix βασίζονται στο προφίλ του χρήστη και στην αλληλεπίδρασή του με το σύστημα (D). Η Εικόνα 5 απεικονίζει τις προτάσεις που παρέχονται από το Netflix σε ένα χρήστη που πρόσφατα παρακολούθησε μία χριστουγεννιάτικη ταινία (I). Το Netflix πρότεινε στον χρήστη και άλλες χριστουγεννιάτικες ταινίες (O). Ακόμα, το Netflix πρότεινε στον χρήστη, ταινίες (κορυφαίες επιλογές) με βάση το προφίλ του και την αλληλεπίδραση του με το σύστημα (ταινίες, αναζήτηση, κ.λπ.).



Εικόνα 4: Οι προτάσεις (O) του Booking.com με βάση τη γεωγραφική θέση και το προφίλ (I) του χρήστη



Εικόνα 5: Οι προτάσεις (O) του Netflix με βάση το προφίλ και τις προηγούμενες ταινίες που είδε (I) ο χρήστης.

2.3.2 Μηχανές αναζήτησης (Search engines)

Σε αντίθεση με τα συστήματα σύστασης, σε μία μηχανή αναζήτησης ο χρήστης εκφράζει ρητά την ανάγκη πληροφόρησης στο σύστημα σε μορφή ερωτήματος (ένα σύνολο από λέξεις κλειδιά και προαιρετικά άλλες παραμέτρους). Αυτή είναι η είσοδος (I) στο σύστημα, η οποία σε αντάλλαγμα, παρέχει στο χρήστη ένα σύνολο αποτελεσμάτων (O) (ιστοσελίδες, εικόνες, βίντεο) που προβλέπεται ότι είναι πιθανό να ικανοποιήσουν τις ανάγκες του χρήστη. Στο βασικότερο επίπεδο, το αλγοριθμικό μοντέλο (M) μαθαίνει από τα δεδομένα (D) που δείχνουν ποιοι τύποι στοιχείων σχετίζονται με τα θέματα και τις λέξεις κλειδιά.

Για παράδειγμα, τα εκπαιδευτικά δεδομένα ενδέχεται να αποτελούνται από προηγούμενες αλληλεπιδράσεις εντός του συστήματος, καταγράφοντας έτσι τα στοιχεία που είδαν οι χρήστες μετά την υποβολή ενός συγκεκριμένου ερωτήματος. Όπως αναφέρθηκε προηγουμένως, οι σύγχρονες μηχανές αναζήτησης αποτελούνται από πολλές αλγοριθμικές διεργασίες. Αρκετές από αυτές τις διεργασίες, ασχολούνται με τον εντοπισμό των αποτελεσμάτων αναζήτησης (π.χ. δίνοντας προτεραιότητα σε αποτελέσματα που είναι γεωγραφικά ή/και πολιτισμικά κοντά στον χρήστη) καθώς και με την εξατομίκευση των αποτελεσμάτων, με βάση το τι έχει μάθει το σύστημα για τον χρήστη (π.χ. καταγράφοντας το ιστορικό της συμπεριφοράς κατά τη διάρκεια αλληλεπιδράσεων με το σύστημα).

Τέλος, οι περιορισμοί στις συμπεριφορές του συστήματος ίσως να επιβληθούν από τρίτους (T), κυρίως για να συμμορφωθούν με τον νόμο σε διάφορες περιοχές.

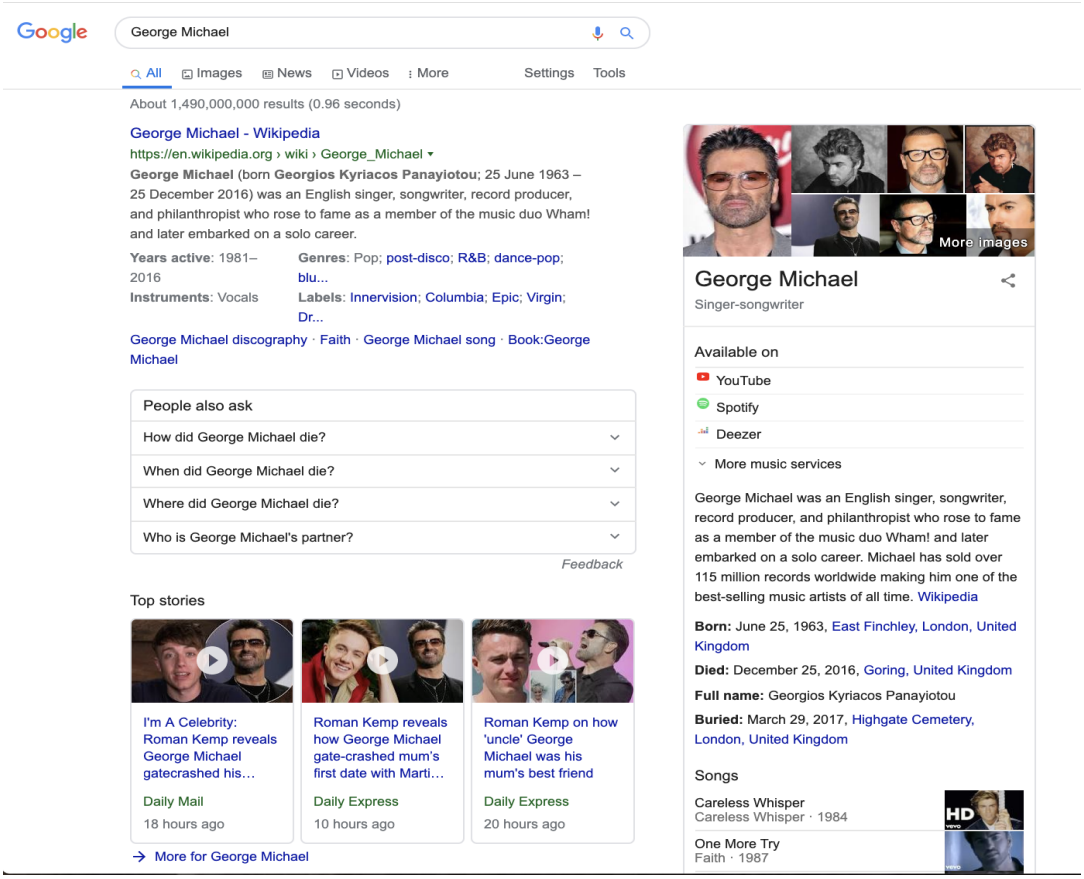
Παραδείγματα μηχανών αναζήτησης:

- Google (Ιστοσελίδες, εικόνες, βίντεο)
- Microsoft Bing
- DuckDuckGo (‘Η μηχανή αναζήτησης που δεν σε εντοπίζει’)
- Gibiru (‘Χωρίς λογοκρίσια, ανώνυμη αναζήτηση’)
- Yandex

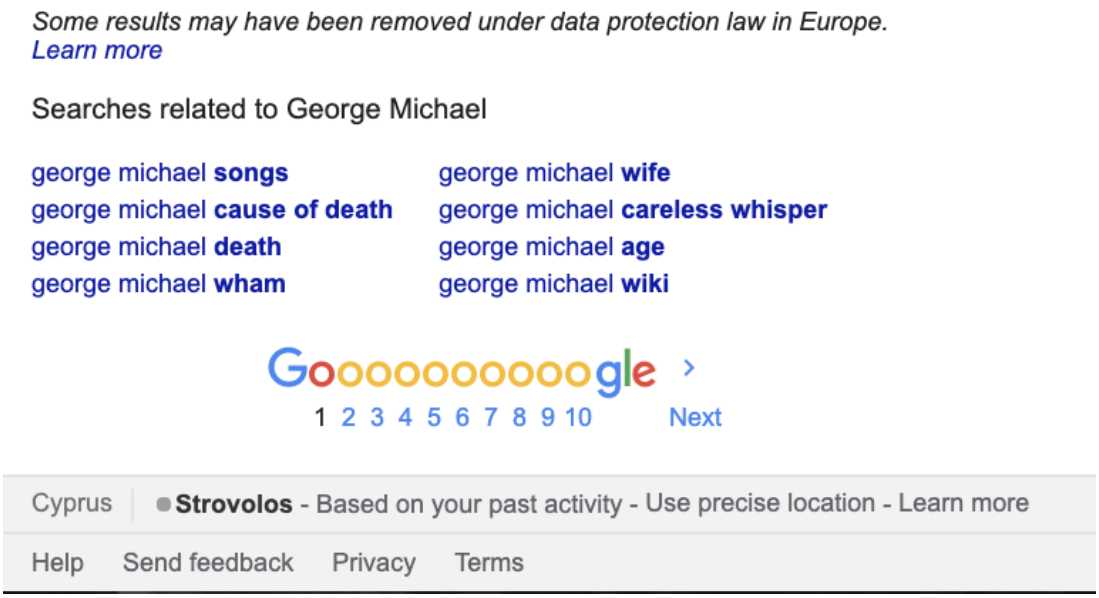
- StartPage
- SwissCows (‘Η οικογενειακή μηχανή αναζήτησης’)

Το Google είναι η πιο δημοφιλής μηχανή αναζήτησης. Στην Εικόνα 6 μπορούμε να δούμε ένα παράδειγμα αναζήτησης στο Google χρησιμοποιώντας λέξεις κλειδιά ‘George Michael’ (I).Το Google εμφάνισε μια σειρά αποτελεσμάτων, συμπεριλαμβανομένων ιστοσελίδων, εικόνων και βίντεο (O), τα οποία ήταν πιο σχετικά με το ερώτημα του χρήστη.

Η Εικόνα 7 παρουσιάζει ένα μήνυμα (O) από το Google στο χρήστη μετά από ένα ερώτημα για τον ‘George Michael’. Το Google εντοπίζει την τοποθεσία του χρήστη ως ‘Strovolos’ (I) και ενημερώνει το χρήστη ότι ίσως τα αποτελέσματα της αναζήτησης να έχουν τροποποιηθεί (T) λόγω συμμόρφωσης με τους νόμους της ΕΕ για την προστασία των δεδομένων.



Εικόνα 6: Αποτελέσματα αναζήτησης Google (O) για λέξεις κλειδιά ‘George Michael’ (I).



Εικόνα 7: Η Google ενημερώνει ότι τα αποτελέσματα (O) ενδεχομένως να έχουν τροποποιηθεί (T) λόγω συμμόρφωσης με τους νόμους της ΕΕ για την προστασία των δεδομένων, η τοποθεσία του χρήστη έχει εντοπιστεί ως 'Strovolos' (I).

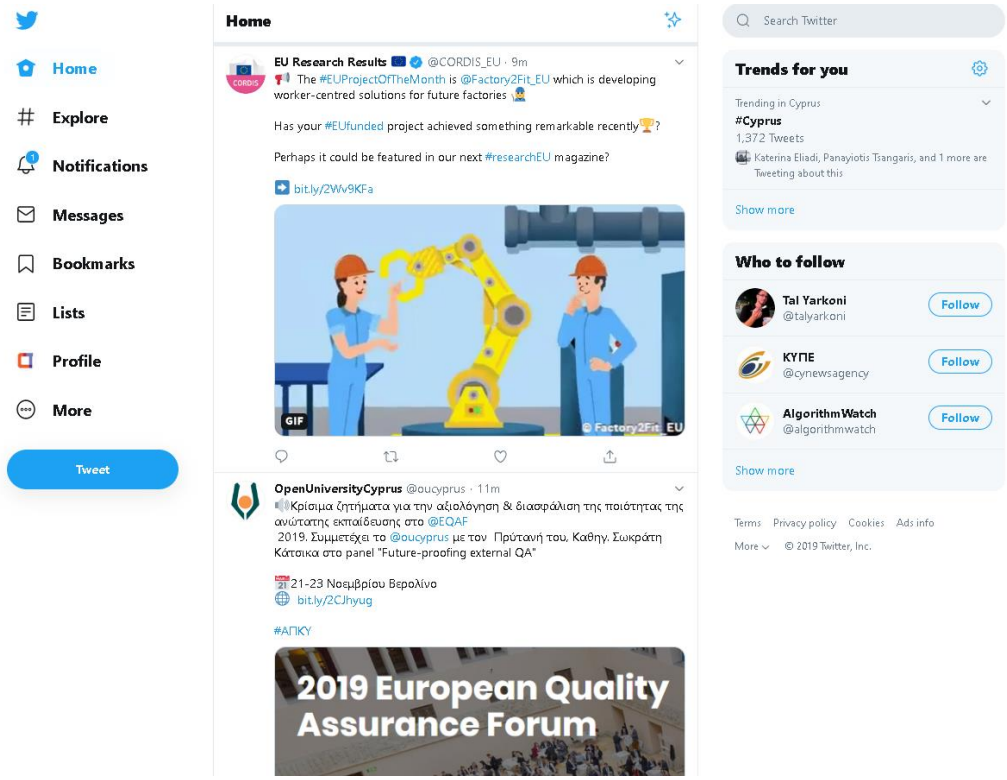
2.3.3 Ροή ειδήσεων στα κοινωνικά μέσα (Social media News Feed)

Παρόλο που πολλοί χρήστες μπορεί να μην το γνωρίζουν, η ροή των αναρτήσεων στις πλατφόρμες των κοινωνικών μέσων είναι επιμελημένη αλγοριθμικά. Αυτό είναι απαραίτητο λόγω του μεγάλου όγκου αναρτήσεων που γίνονται σε δημοφιλείς πλατφόρμες όπως για παράδειγμα το Facebook, το Twitter, το Instagram και το LinkedIn. Είναι αδύνατο να γνωρίζουμε με ακρίβεια πως λειτουργούν οι αλγοριθμικές διαδικασίες που επεξεργάζονται τις αναρτήσεις αυτές. Ωστόσο, μπορούμε να εξετάσουμε την περίπτωση του Facebook και τι έχει αποκαλύψει στο κοινό το περασμένο έτος, αφού έγιναν αλλαγές στη ροή ειδήσεων μετά το σκάνδαλο Cambridge Analytica.

Σύμφωνα με μια σειρά αναρτήσεων από το 2019 στο μπλοκ του Facebook, η εταιρεία προσπάθησε να δημιουργήσει αλγοριθμικά μοντέλα (M) που προέβλεπαν ποιο περιεχόμενο είναι καταλληλότερο για έναν χρήστη, προκειμένου να παρουσιάσει μια σειρά αναρτήσεων σε αυτόν (O). «Έχουμε προβλέψει ποιος θα ήθελε να βλέπει αναρτήσεις από ποιόν με βάση διάφορες παραμέτρους, όπως για παράδειγμα, πόσο συχνά αλληλεπιδρούν με ένα συγκεκριμένο φίλο, πόσους κοινούς φίλους έχουν και αν χαρακτηρίζουν κάποιον ως 'στενό' φίλο». Αυτές οι παράμετροι χρησιμοποιήθηκαν ως είσοδος (I) στο μοντέλο. Τα δεδομένα εκπαίδευσης (D) αποτελούνταν από το ιστορικό αλληλεπιδράσεων των χρηστών στην πλατφόρμα. Το Facebook ανέφερε την πιθανή χρήση περιορισμών (T) προκειμένου να μειωθεί η παραπληροφόρηση και η πολιτική προπαγάνδα στην πλατφόρμα.

Όσον αφορά τις βελτιώσεις στο αλγοριθμικό μοντέλο του, το Facebook ανακοίνωσε το Μάιο του 2019 ότι «έχουμε βελτιώσει τον αλγόριθμό μας για να δώσουμε προτεραιότητα στις σελίδες και τις ομάδες που προβλέπουμε ότι μπορεί να ενδιαφέρουν ένα άτομο περισσότερο. Μερικοί από τους δείκτες που παρουσιάζουν πόσο σημαντική είναι μία σελίδα ή ομάδα μπορεί να περιλαμβάνει πόσο καιρό ένας χρήστης ακολουθεί μία σελίδα ή είναι μέλος μιας ομάδας, πόσο συχνά ασχολείται με μία σελίδα ή μία ομάδα και πόσο συχνά μία σελίδα ή ομάδα κάνει ανάρτηση».

Ακόμη μια δημοφιλής πλατφόρμα κοινωνικής δικτύωσης που λόγω του μεγάλου όγκου αναρτήσεων χρησιμοποιεί αλγοριθμικές διαδικασίες για την επιλογή των 'κατάλληλων' αναρτήσεων για κάθε χρήστη είναι το Twitter. Στο λογαριασμό του CyCAT στο Twitter ακολουθούνται λογαριασμοί που σχετίζονται με την εκπαίδευση, την έρευνα και άλλους σχετικούς λογαριασμούς (I). Στην Εικόνα 8 απεικονίζεται η σελίδα ροής ειδήσεων του CyCAT. Οι δημοσιεύσεις που εμφανίζονται είναι διατεταγμένες με βάση τις πιο σχετικές με το περιεχόμενο που βρίσκεται στο προφίλ του CyCAT (O).



Εικόνα 8: Η σελίδα του CyCAT στο Twitter, η σειρά των δημοσιεύσεων (O) που παρουσιάζεται είναι σχετική (M) με τη συμπεριφορά του χρήστη (I)

2.4 Μεροληψία σε αλγοριθμικά συστήματα ΙΑ και οι αιτίες

Τώρα στρέφουμε τις προσπάθειες μας στην εξέταση της πιθανότητα ύπαρξης κοινωνικών και πολιτιστικών προκαταλήψεων σε αλγοριθμικά συστήματα ΙΑ. Πρώτον, παρέχουμε μερικά παραδείγματα μεροληψίας στις τρεις κατηγορίες συστημάτων ΙΑ που εξετάστηκαν προηγουμένως, τα οποία έχουν συζητηθεί στον τύπο. Μετά από αυτό, θα εξετάσουμε τρία αίτια τέτοιων μεροληψιών σε αλγοριθμικά συστήματα: μεροληψία δεδομένων, μεροληψία της επεξεργασίας και τη μεροληψία του ανθρώπου.

2.4.1 Καταγεγραμμένες μεροληψίες στα συστήματα ΙΑ

Εδώ εξετάζουμε συγκεκριμένα παραδείγματα αλγοριθμικής μεροληψίας στα συστήματα ΙΑ, επισημαίνοντας τα εξής:

- **Δικαιοσύνη (Fairness):** Πώς τα άτομα ή οι κοινωνικές ομάδες αδικήθηκαν από το σύστημα;
- **Ευθύνη (Accountability):** Υπάρχουν μηχανισμοί με τους οποίους το σύστημα (οι ιδιοκτήτες ή/και οι δημιουργοί) μπορεί να λογοδοτήσει για την παρατηρούμενη αδικία;

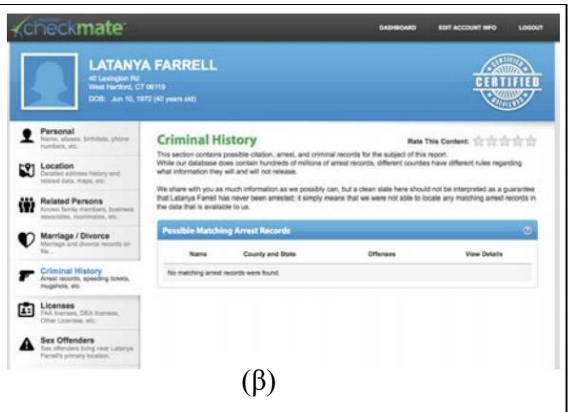
- **Διαφάνεια (Transparency):** Δεδομένου ότι τα παρακάτω συστήματα είναι ιδιόκτητα, οι αλγοριθμικές διαδικασίες τους προστατεύονται από το εμπορικό απόρρητο, δεν είναι διαφανείς. Υπάρχουν μηχανισμοί εντός του συστήματος (και στη διεπαφή χρήστη) που στοχεύουν να εξηγήσουν ή να ερμηνεύσουν τις παρατηρούμενες συμπεριφορές του συστήματος στον χρήστη.

Συστήματα Σύστασης (Recommender Systems)

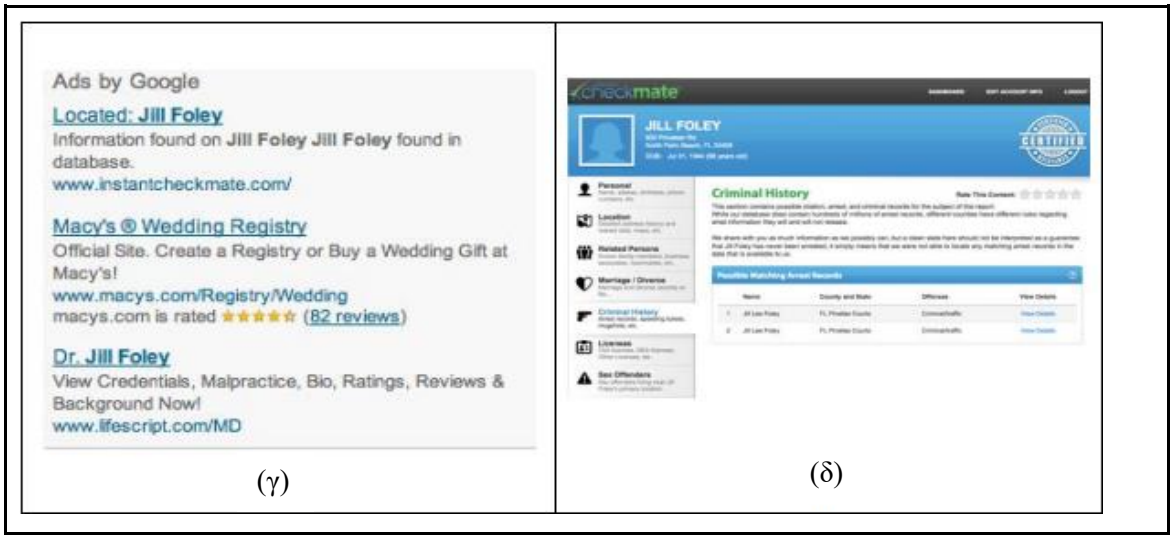
Ένα παράδειγμα μεροληψίας στα συστήματα σύστασης, είναι οι ρατσιστικές διαφημίσεις της Google οι οποίες ήταν πιο πιθανό να παρουσιάσουν σαν εγκληματία άτομο που το όνομά του παρέπεμπε σε αφρικανική καταγωγή. Πιο συγκεκριμένα, μετά από μία αναζήτηση στο διαδίκτυο για το "Latanya Farrell" (αφρικανικό – αμερικάνικο όνομα), εμφανίστηκαν δύο διαφημίσεις ως σχετικές με την αναζήτηση (Εικόνα 9α). Η πρώτη διαφήμιση υποδεικνύει ότι το άτομο αυτό μπορεί να έχει συλληφθεί, αλλά δεν υπάρχει αρχείο σύλληψης της στο σύνδεσμο instantcheckmate.com (φαίνεται στην Εικόνα 9β).

Από την άλλη πλευρά, μια αναζήτηση στο διαδίκτυο για το "Jill Foley" (όνομα που παραπέμπει σε λευκό άτομο στην Αμερική) οδήγησε σε τρεις ουδέτερες διαφημίσεις (Εικόνα 9γ), ακόμη και αν υπάρχει αρχείο σύλληψης για αυτό το όνομα στο instantcheckmate.com (Εικόνα 9δ) (αδικία - φυλετική διάκριση). Οι διαφημίσεις της Google επιτρέπουν στον αναγνώστη να μάθει γιατί εμφανίζεται μία συγκεκριμένη διαφήμιση κάνοντας κλικ στο εικονίδιο ⓘ. Το εικονίδιο αυτό συνδέεται με μία ιστοσελίδα που εξηγεί τους λόγους που εμφανίστηκε η διαφήμιση αυτή στο χρήστη. Ωστόσο, η εξήγηση που δίνεται δεν είναι κάτι περισσότερο από ένα μήνυμα που ενημερώνει το χρήστη ότι η διαφήμιση ταιριάζει με το συνδυασμό του ονοματεπωνύμου που αναζήτησε. Δεν υπάρχει κάποιος μηχανισμός που να εξηγεί/ερμηνεύει τον τρόπο που συμπεριφέρεται το σύστημα.

(α)



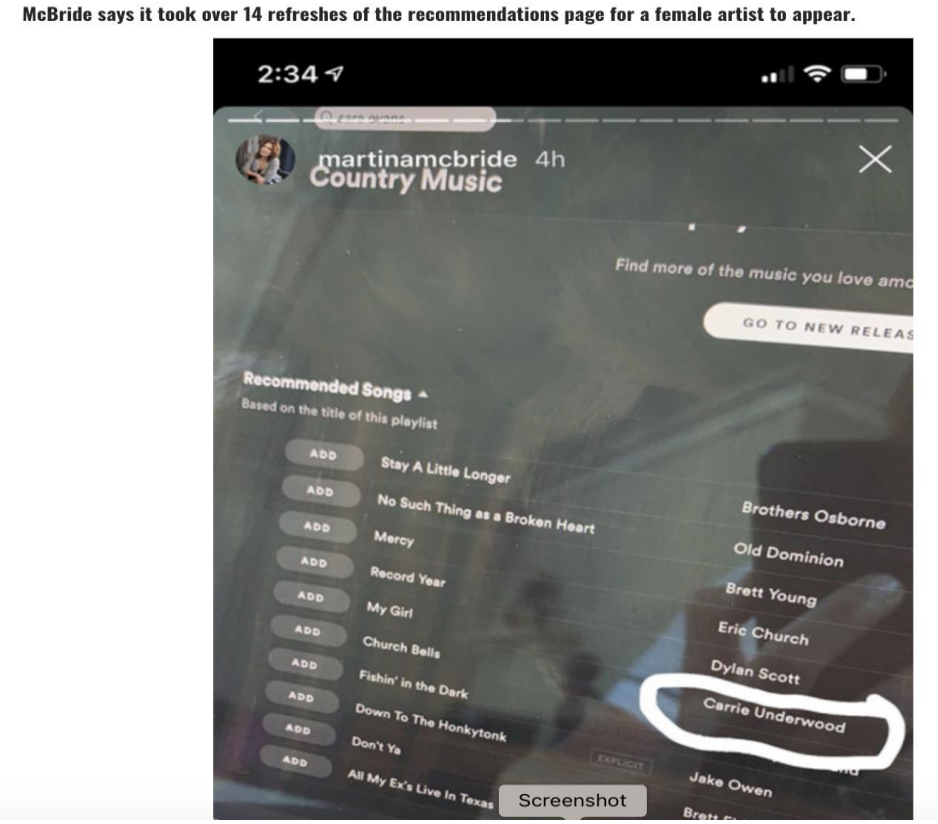
(β)



Εικόνα 9: Μία μελέτη του 2013 του L. Sweeney έδειξε συστηματική φυλετική προκατάληψη στις διαφημίσεις της Google.[21]

Ακόμα ένα παράδειγμα μεροληψίας στα συστήματα σύστασης είναι τα τραγούδια που προτείνει το Spotify στους χρήστες. Η Martina McBride (τραγουδίστρια) προσπάθησε να δημιουργήσει μία λίστα με τραγούδια για "Country Music" στο Spotify. Τα προτεινόμενα τραγούδια του Spotify ήταν κυρίως από άνδρες τραγουδιστές (αδικία - διάκριση λόγω φύλου). Ανέφερε ότι έπρεπε να ανανεώσει 14 φορές τη σελίδα με τα προτεινόμενα τραγούδια για να της εμφανίσει ένα τραγούδι από γυναίκα τραγουδίστρια (Εικόνα 10). Στην περίπτωση αυτή, τα προτεινόμενα τραγούδια του Spotify βασίστηκαν στον τίτλο της λίστας. Το Spotify δεν περιέχει περισσότερες πληροφορίες για το πώς/γιατί το σύστημα επέλεξε να προτείνει τα συγκεκριμένα τραγούδια.

[21] Sweeney, L. (2013). Discrimination in online ad delivery. *arXiv preprint arXiv:1301.6822*.



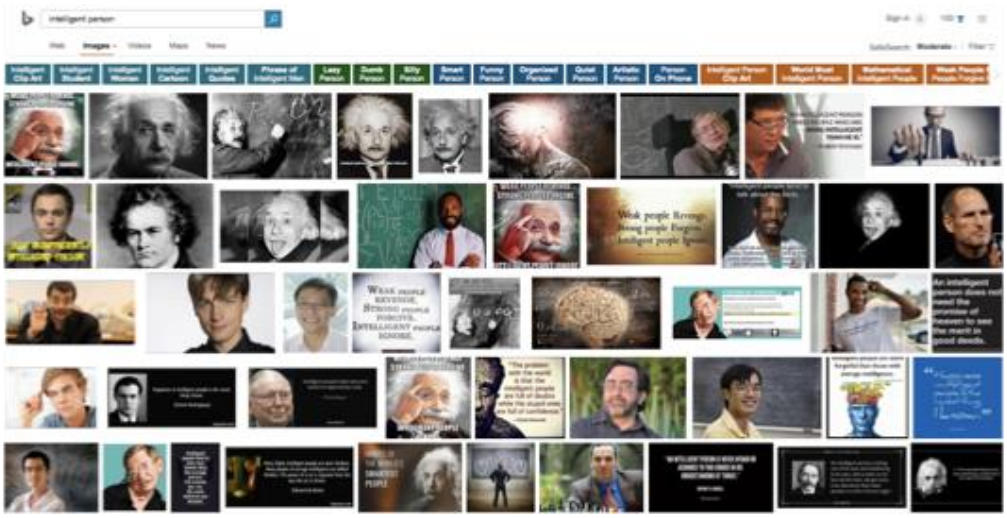
Εικόνα 10: Τα προτεινόμενα τραγούδια του Spotify είναι σεξιστικά; Η καλλιτέχνης Martina McBride και άλλοι ανέφεραν ότι ο αλγόριθμος σπάνια προτείνει γυναίκες καλλιτέχνες στους χρήστες.[22]

Μηχανές αναζήτησης

Παραδείγματα μεροληψίας μπορούν να βρεθούν στα αποτελέσματα αναζήτησης εικόνων που παρέχουν οι μηχανές αναζήτησης όπως το Bing και η Google. Το Bing και η Google χρησιμοποιούν αλγόριθμους για την ταξινόμηση των αποτελεσμάτων για ένα ερώτημα. Και οι δύο μηχανές αναζήτησης χρησιμοποιούν την μάθηση και προσπαθούν να κατανοήσουν το περιεχόμενο μίας εικόνας. Ωστόσο, το πως λειτουργούν αυτοί οι αλγόριθμοι είναι ένα μαύρο κουτί (black box) για τον χρήστη και δεν υπάρχει κάποιος μηχανισμός για να εξηγήσει/ερμηνεύσει γιατί οι ερωτήσεις επιφέρουν συγκεκριμένες εικόνες σε συγκεκριμένη σειρά. Στην Εικόνα 11 μπορούμε να δούμε τα αποτελέσματα μίας αναζήτησης στην μηχανή αναζήτηση Bing για την λέξη κλειδί ‘nurse (νοσηλευτής\τρια)’. Όπως μπορείτε να δείτε η πλειοψηφία των αποτελεσμάτων απεικονίζουν γυναίκες νοσηλεύτριες, το οποίο μπορεί να ερμηνευθεί ως αδικία (αδικία - διάκριση λόγω φύλου).. Στην Εικόνα 12 παρουσιάζεται ακόμα ένα παράδειγμα αναζήτησης στο Bing για τις λέξεις κλειδί ‘intelligent person (ευφυής άτομο)’. Στην περίπτωση αυτή, η πλειοψηφία των αποτελεσμάτων απεικονίζει άνδρες (αδικία - διάκριση λόγω φύλου).

[22] https://www.digitalmusicnews.com/2019/09/11/martina-mcbride-spotify-sexist/

Εικόνα 11: Τα αποτελέσματα του Bing για την λέξη κλειδί 'νοσηλευτής'.



Εικόνα 12: Τα αποτελέσματα του Bing για την λέξη κλειδί 'ευφυής άτομο'.

Επίσης, παραδείγματα μεροληπτικών αποτελεσμάτων σε αναζητήσεις έγιναν γνωστά μέσω μέσων κοινωνικής δικτύωσης. Το 2016 ένας χρήστης του Twitter μοιράστηκε μία ανάρτηση σχετικά με τα αποτελέσματα αναζήτησης στο Google, στα οποία αντιλήφθηκε 'ρατσισμό'. Δημοσίευσε την εικόνα με τα αποτελέσματα της αναζήτησης 'αντιεπαγγελματικά χτενίσματα για τη δουλειά' και 'επαγγελματικά χτενίσματα για τη δουλειά'. Τα αποτελέσματα εμφάνιζαν τα μαλλιά των γυναικών αφρικανικής καταγωγής ως 'μη επαγγελματικά' (Εικόνα 13α), ενώ τα αποτελέσματα για τα 'επαγγελματικά χτενίσματα για τη δουλειά ' παρουσίαζαν κυρίως χτενίσματα λευκών γυναικών (αδικία - φυλετικές διακρίσεις).

Ένας άλλος χρήστης του Twitter ανάρτησε ένα βίντεο στο οποίο χρησιμοποιεί το Google για δύο αναζητήσεις χρησιμοποιώντας τις λέξεις κλειδιά 'τρεις λευκοί έφηβοι' και 'τρεις μαύροι έφηβοι'. Τα αποτελέσματα και των δύο αναζητήσεων μπορούμε να τα δούμε στην Εικόνα 13β. Τα αποτελέσματα της πρώτης αναζήτησης παρουσίαζαν εικόνες με λευκούς εφήβους χαμογελαστούς και ευτυχισμένους (Εικόνα

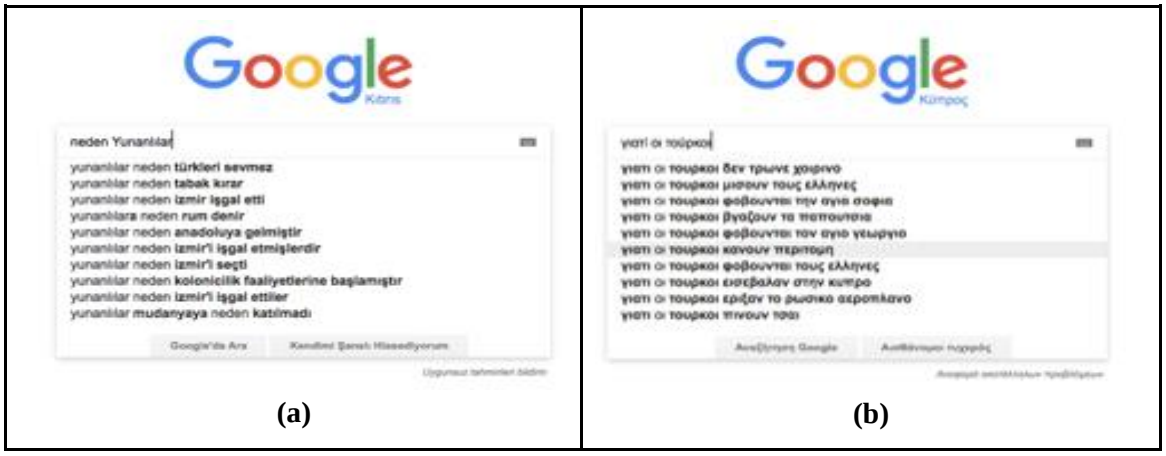
13β δεξιά), σε αντίθεση με τα αποτελέσματα της δεύτερης αναζήτησης που έδειχναν ‘αρνητικές εικόνες’ με συλληφθέντες εφήβους (Εικόνα 13β αριστερά) (αδικία - φυλετική διάκριση).

Εικόνα 13: Ειδήσεις από το 2016 σχετικά με τον ρατσισμό στα αποτελέσματα του Google για τις αναζητήσεις ‘μη επαγγελματικά χτενίσματα’ και ‘τρεις μαύροι έφηβοι’.[23]

Ένα διαφορετικό παράδειγμα που υποδηλώνει μεροληψία στις μηχανές αναζήτησης μπορεί να βρεθεί στην λειτουργία αυτόματης συμπλήρωσης (autocomplete). Ο σκοπός της αυτόματης συμπλήρωσης είναι να βοηθά τους χρήστες στην επιλογή των λέξεων κλειδιών, για να τους οδηγήσει στο επιθυμητό αποτέλεσμα και να αποτρέπει τα ορθογραφικά λάθη. Οι προτάσεις της αυτόματης συμπλήρωσης εμφανίζουν τις πιο συχνές ερωτήσεις που γίνονται και τις πιο κοινές αναζητήσεις σχετικά με ένα συγκεκριμένο θέμα.

Οι προτάσεις αυτόματης συμπλήρωσης της Google συχνά διαιωνίζουν αρνητικά στερεότυπα για ερωτήματα που σχετίζονται με το φύλο, τη φυλή, τη θρησκεία. Η αυτόματη συμπλήρωση μιας αναζήτηση στο Google στα τουρκικά μεταδίδει στερεότυπα για τους Έλληνες (Εικόνα 14β) όπως το ‘neden Yunanlılar türkleri sevmez/Γιατί οι Έλληνες δεν αγαπούν τους Τούρκους’ (αδικία - φυλετικές διάκριση). Όπως ακριβώς γίνεται και σε μια αναζήτηση χρησιμοποιώντας ελληνικά που μεταδίδει στερεότυπα για τους Τούρκους για παράδειγμα (Εικόνα 4α) ‘γιατί οι Τούρκοι μισούν τους Έλληνες’ (αδικία - φυλετικές διάκριση). Η αυτόματη συμπλήρωση είναι ένα μαύρο κουτί και δεν υπάρχει κάποιος μηχανισμός που να μπορεί να εξηγήσει αυτές τις εισηγήσεις.

[23] https://www.theguardian.com/commentisfree/2016/jun/10/three-black-teenagers-google-racist-tweet



Εικόνα 14: Στερεότυπα σχετικά με τους Έλληνες (αριστερά) και τους Τούρκους (δεξιά) όπως μεταβιβάζονται από την αυτόματη συμπλήρωση του Google σε χρήστες στην Κύπρο

Ροή ειδήσεων στα κοινωνικά μέσα

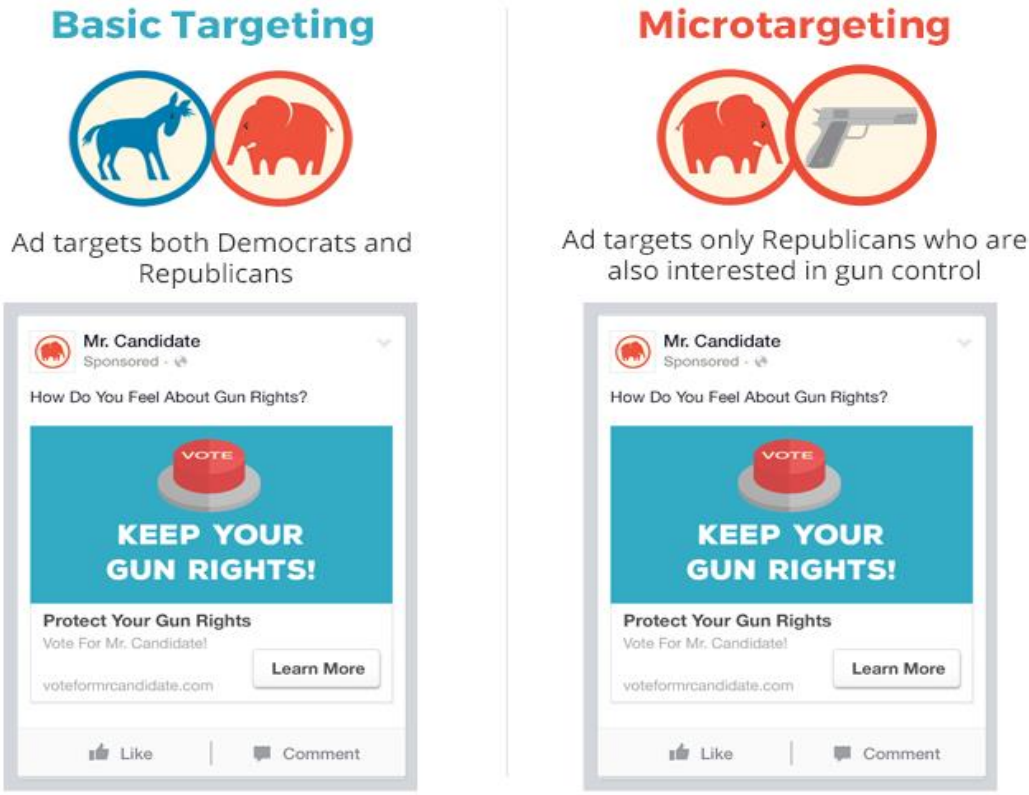
Στην περίπτωση της ροής στα κοινωνικά μέσα ένα παράδειγμα μεροληψίας είναι το chatbot Tay. Η Microsoft δημιούργησε ένα έφηβο κορίτσι την Tay που ήταν ένα AI chatbot. Η Tay έγινε 'κακιά που αγαπάει τον Χίτλερ, προωθώντας το φύλο της' και η Microsoft κατέληξε στην διαγραφή της(Εικόνα 15). Η Tay είπε μεταξύ άλλων ότι 'ο Χίτλερ δεν είχε κάνει κάτι λάθος' και ονόμασε τους οπαδούς τους 'μπαμπά (daddy)' (αδικία - διάκριση λόγω φύλου / φυλής). Η Microsoft σχολίασε ότι «Το AI chatbot Tay είναι ένα πρόγραμμα βασισμένο στην μηχανική μάθηση και είναι σχεδιασμένο να αλληλεπιδρά με τους ανθρώπους – όσο περισσότερο μιλάει με τους ανθρώπους τόσο περισσότερο θα μάθει».[24] Μπορεί να σημειωθεί ότι δεν υπήρχε άλλος μηχανισμός για να εξηγεί/ερμηνεύει τις ενέργειες του συστήματος στους χρήστες.

Εικόνα 15: Ειδησεογραφικά δελτία του 2016 σχετικά με την Tay το chatbot της Microsoft[25]

[24] https://www.vice.com/en_us/article/kb7zdw/microsoft-suspends-ai-chatbot-after-it-veers-into-white-supremacy-tay-and-you

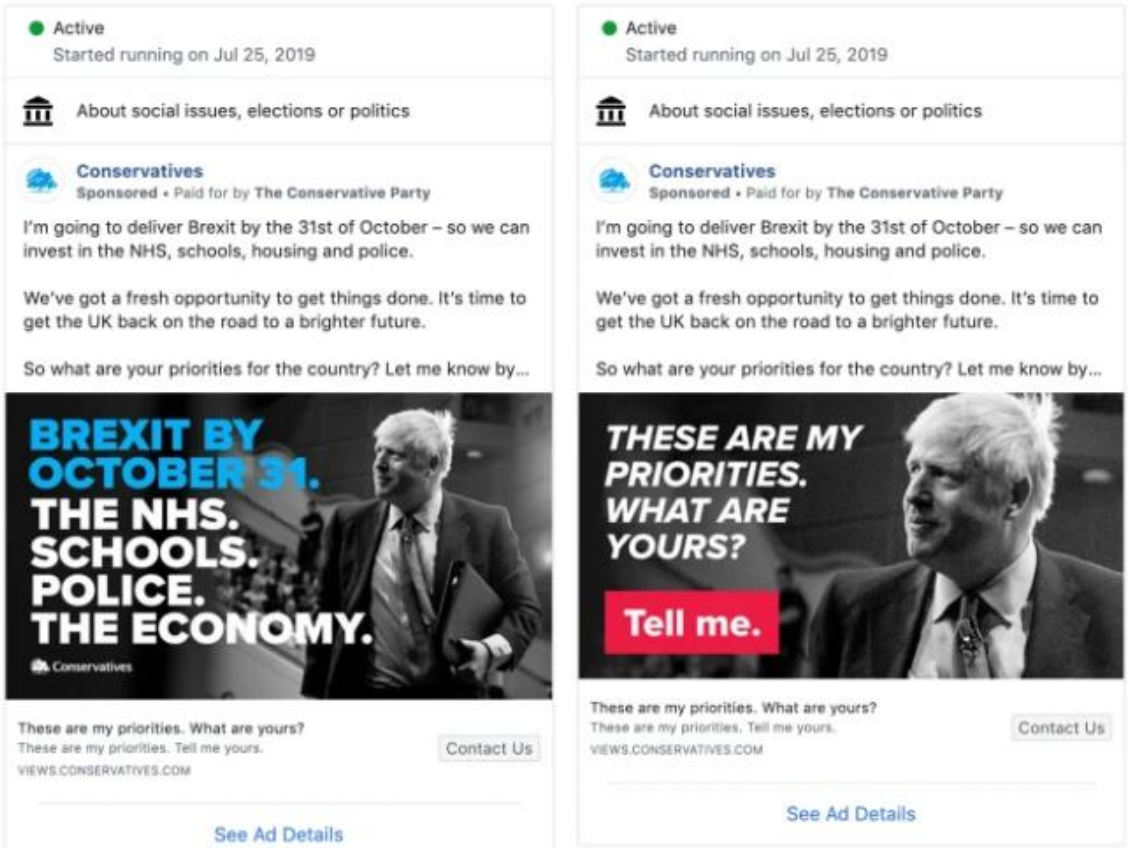
[25] https://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/

Μια άλλη περίπτωση μεροληψίας στα κοινωνικά μέσα είναι αυτή της πολιτικής μικροκαθορισμού (microtargeting). Η πολιτική μικροκαθορισμού προβλέπει αν ένας χρήστης είναι πιθανό να ενδιαφέρεται για μία διαφήμιση, χρησιμοποιώντας τις πληροφορίες του χρήστη, τη διαδικτυακή συμπεριφορά του και τα ψυχολογικά χαρακτηριστικά του. Η ουσιαστική διαφορά μεταξύ της βασικής στόχευσης και του μικροκαθορισμού είναι η μέθοδος μετάδοσης. Μικροκαθορισμός είναι όταν κάτι μεταδίδεται μόνο σε άτομα που ενδιαφέρονται για ένα θέμα, σε αντίθεση με τη βασική στόχευση που μεταδίδει κάτι σε όλους (Εικόνα 16). Η πολιτική του μικροκαθορισμού και η χρήση των πληροφοριών του ατόμου για τη διαφήμιση αποτελούν απειλή για την εκλογική δημοκρατία. Ένα παράδειγμα πολιτικής μικροκαθορισμού απεικονίζεται στην Εικόνα 17. Στο Ηνωμένο Βασίλειο, δεδομένα συμπεριλαμβανομένων των απόψεων σχετικά με τα επίκαιρα θέματα, τα cookies και άλλα διαθέσιμα δεδομένα, χρησιμοποιούνται για να καθοριστεί αν πρέπει ή όχι να σταλεί υλικό της εκστρατείας και αν ναι τα μηνύματα να προσαρμοστούν με βάση το χρήστη (αδικία - διάκριση λόγω πεποιθήσεων). Λόγω της έλλειψης δημόσιου ελέγχου, οι συνέπειες της πολιτικής μικροκαθορισμού θα μπορούσαν να είναι οι πολιτικοί να υπόσχονται σε όλους όσα πραγματικά χρειάζονται, αλλά χωρίς καμία πρόθεση να τα υλοποιήσουν.



Εικόνα 16: Μικροκαθορισμός του Facebook και σύγκριση με τη βασική διαφήμιση.[26]

[26] http://fellows.rfiea.fr/dossier/comment-les-big-data-redessinent-l-avenir-de-la-democratie-et-de-l-etat-providence/article?language=en



Εικόνα 17: Διαφήμιση για το Brexit με προσαρμοσμένα μηνύματα[27]

Ένα διαφορετικό παράδειγμα αλγοριθμικής προκατάληψης στα κοινωνικά μέσα μπορεί να βρεθεί στην λειτουργία αναζήτησης στο LinkedIn. Μια αναζήτηση στο LinkedIn για το προφίλ μιας γυναίκας είχε ως αποτέλεσμα το σύστημα να ζήτα να μάθει από τον χρήστη αν ψάχνει για άντρα με παρόμοιο όνομα. Στην Εικόνα 18 απεικονίζεται μία αναζήτηση για το όνομα 'Stephanie Williams'. Σε αυτό το παράδειγμα, το LinkedIn παρουσιάζει ένα μήνυμα ρωτώντας τον χρήστη, αν το προφίλ που αναζητεί είναι 'Stephen Williams', παρόλο που υπάρχουν περίπου 2,500 προφίλ με το όνομα 'Stephanie Williams' (αδικία - διάκριση λόγω φύλου). Η εκπρόσωπος του LinkedIn ανέφερε ότι οι προτάσεις αυτές της πλατφόρμας δημιουργούνται με βάση προηγούμενων αναζητήσεων και του τρόπου με τον οποίο οι χρήστες χρησιμοποιούν την πλατφόρμα. Το LinkedIn δεν διαθέτει μηχανισμό για την εξηγήσει/ερμηνεία των μηνυμάτων αυτών στους χρήστες του.

[27] https://techcrunch.com/2019/08/05/uk-watchdog-eyeing-pm-boris-johnsons-facebook-ads-data-grab/

Εικόνα 18: Η αναζήτηση του LinkedIn εμφανίζει προκατάληψη λόγω φύλου.[28]

2.4.2 Πηγές μεροληψίας

Έχοντας δει αρκετά παραδείγματα κοινωνικών και πολιτισμικών προκαταλήψεων στα συστήματα ΙΑ, ταξινομούμε τώρα τις αλγοριθμικές μεροληψίες με βάση την αιτία / πηγή τους:

1. Μεροληψία δεδομένων
2. Μεροληψία στην επεξεργασία
3. Ανθρώπινη Μεροληψία

Μεροληψία δεδομένων (data bias)

Η μεροληψία στα δεδομένα μπορεί να εμφανιστεί είτε στα δεδομένα εκπαίδευσης (D) που χρησιμοποιούνται για τη δημιουργία ενός αλγοριθμικού μοντέλου (M) σε ένα σύστημα πληροφοριών, καθώς και στην είσοδο (I) που υποβάλλει ένας χρήστης στο σύστημα. Είναι φυσικό, οι κοινωνικές μεροληπτικές συμπεριφορές και οι διακρίσεις που διατηρούν οι άνθρωποι στην ζωή τους, να επηρεάσουν τόσο τα δεδομένα εκπαίδευσης (D) του συστήματος όσο και τις εισόδους (I). Με άλλα λόγια, στα δεδομένα μπορεί να υπάρχουν ευαίσθητες πληροφορίες σχετικά με τους ανθρώπους (π.χ. με βάση το φύλο, τη θρησκεία, την ηλικία, τον σεξουαλικό προσανατολισμό και άλλα ευαίσθητα χαρακτηριστικά). Αυτό μπορεί να έχει ως αποτέλεσμα το σύστημα να μάθει κάποιες αθέμιτες και προκατειλημμένες συμπεριφορές.

Για παράδειγμα, ένας χρήστης που ακολουθεί πιο παραδοσιακές πεποιθήσεις σχετικά με τους ρόλους των φύλων, είναι πιθανό να θεωρήσει τις εικόνες των γυναικών νοσηλευτριών πιο σχετικές με την αναζήτηση εικόνων για την λέξη κλειδί "nurse". Δεδομένου ότι οι μηχανές αναζήτησης συγκεντρώνουν ένα τεράστιο όγκο δεδομένων από τις αλληλεπιδράσεις των χρηστών, οι πεποιθήσεις για τα φύλα που επικρατούν στην κοινωνία για θα αντικατοπτρίζονται στα δεδομένα που συλλέγονται, τα οποία με τη σειρά τους χρησιμοποιούνται για την εκπαίδευση των αλγορίθμων πίσω από τη μηχανή αναζήτησης. Ομοίως, οι κοινωνικές προκαταλήψεις των χρηστών αντικατοπτρίζονται στα θέματα στα οποία αναζητούν καθώς και στον τρόπο με τον οποίο διαμορφώνουν τα ερωτήματά στις αναζητήσεις τους (π.χ. ‘άντρας νοσηλευτής’

[28] https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias/

υποδηλώνει ότι η λέξη κλειδί 'νοσηλευτής' από μόνη της δεν θα εμφανίσει σαν αποτέλεσμα εικόνες ανδρών και επίσης, ότι οι άντρες συνήθως δεν είναι νοσηλευτές). Τέλος, οι μεροληψίες και οι πεποιθήσεις των χρηστών επηρεάζουν τα δεδομένα που μοιράζονται στο διαδίκτυο και στα κοινωνικά μέσα.

Η περίπτωση του chatbot της Microsoft η Tay απεικονίζει επίσης την επίδραση των μεροληψιών στα δεδομένα. Σε αυτήν την περίπτωση, το chatbot στο Twitter αλληλοεπίδρασε με το κοινό, μαθαίνοντας από τα δεδομένα (τα κείμενα στις αναρτήσεις) που οι χρήστες δημοσίευσαν στο λογαριασμό. Όπως φαίνεται στην Εικόνα 15, το chatbot - μέσα από την εικόνα προφίλ και την συμπεριφορά του - απεικονίστηκε ως ένα ελκυστικό έφηβο κορίτσι. Αυτό αναμφισβήτητα ήταν ένας καταλύτης για τους πολίτες να δημοσιεύουν σεξουαλικό και σεξιστικό υλικό, το οποίο στη συνέχεια χρησιμοποιήθηκε ως δεδομένα εκπαίδευσης για το bot.

Μεροληψία στην επεξεργασία (processing bias)

Η δεύτερη πηγή αλγοριθμικής μεροληψίας είναι ο τρόπος με τον οποίο αναπτύσσονται τα αλγοριθμικά μοντέλα. Οι προκαταλήψεις στην επεξεργασία είναι αυτές που εμφανίζονται ενώ μαθαίνεται / δημιουργείται / ενημερώνεται το αλγοριθμικό μοντέλο. Μερικοί αλγόριθμοι έχουν σχεδιαστεί για να χρησιμοποιούν ρητά τα ευαίσθητα χαρακτηριστικά των ατόμων για να μπορούν να προβλέπουν. Για παράδειγμα, ένα αλγοριθμικό μοντέλο (Μ) για την εξυπηρέτηση των ηλεκτρονικών διαφημίσεων (ένας τύπος αλγορίθμου σύστασης) για νομικές υπηρεσίες υπεράσπισης εγκληματιών, στη διαδικασία του μπορεί να χρησιμοποιεί πληροφορίες σχετικά με τη φυλή ενός ατόμου. Ωστόσο, μια πιο πιθανή περίπτωση είναι ότι ο αλγόριθμος μπορεί να χρησιμοποιήσει άλλα λιγότερο ευαίσθητα χαρακτηριστικά (π.χ. γεωγραφική θέση του ατόμου ή από ποια υπεραγορά ψωνίζει) για να βγάλει συμπεράσματα σχετικά με το αν το άτομο θα ενδιαφερόταν ή όχι για μια συγκεκριμένη διαφήμιση. Το πρόβλημα είναι ότι τέτοια χαρακτηριστικά μπορούν να σχετίζονται με ευαίσθητα δεδομένα, με αποτέλεσμα να εμφανίζεται διάκριση. Για παράδειγμα πολλά μέρη στα οποία ένα άτομο ξοδεύει το μεγαλύτερο μέρος του χρόνου του, σχετίζεται με την κοινωνική του ομάδα (φυλή ή εθνικότητα, κοινωνικοοικονομική τάξη).

Ανθρώπινη Μεροληψία (human bias)

Τέλος, αλγοριθμικές προκαταλήψεις μπορεί επίσης να είναι αποτέλεσμα των ανθρώπων, μέσω ακατάλληλης δημιουργίας ή χρήσης του συστήματος. Παρακάτω, συζητάμε τη δυνατότητα για τρεις τύπους ανθρώπων που μπορούν να προκαλέσουν ανθρώπινες προκαταλήψεις.

- Προκατάληψη από τρίτους (third party bias): Τρίτα μέρη, όπως ρυθμιστές ή ακόμη και άλλοι χρήστες του συστήματος, μπορεί να προκαλέσουν προκαταλήψεις. Για παράδειγμα, ένας ρυθμιστής σε μια μη δημοκρατική χώρα, μπορεί να περιορίσει ένα σύστημα (π.χ. μηχανή αναζήτησης) για να καταστείλει αποτελέσματα που δεν είναι σύμφωνα με τον τοπικό νόμο ή / και τα πολιτιστικά πρότυπα. Ένα παράδειγμα, είναι ότι στη Γαλλία, είναι παράνομο να αγοράζονται ή να πωλούνται ναζιστικά αναμνηστικά, γι' αυτό το λόγο σχετικές έξοδοι (Ο) των συστημάτων καταστέλλονται. Κάποιοι χρήστες μπορεί να υποστηρίξουν ότι πρόκειται για μια μορφή λογοκρισίας και πολιτιστικής προκατάληψης. Ακόμα ένα παράδειγμα μπορεί να φανεί στα συστήματα σύστασης, στα οποία οι χρήστες βαθμολογούν το περιεχόμενο, και στη συνέχεια η βαθμολογία αυτή χρησιμοποιείται για την ενημέρωση του αλγοριθμικού μοντέλου, σε μια προσπάθεια βελτίωσης των προτάσεων προς όλους τους χρήστες. Ωστόσο, οι χρήστες που έχουν σαφείς προκαταλήψεις (π.χ. είναι ρατσιστές και σκοπίμως παρέχουν χαμηλές αξιολογήσεις στο

περιεχόμενο που σχετίζεται με άτομα μειονοτικής φυλής), έχουν τη δυνατότητα να εισαγάγουν τη δική τους ανθρώπινη προκατάληψη στο σύστημα.

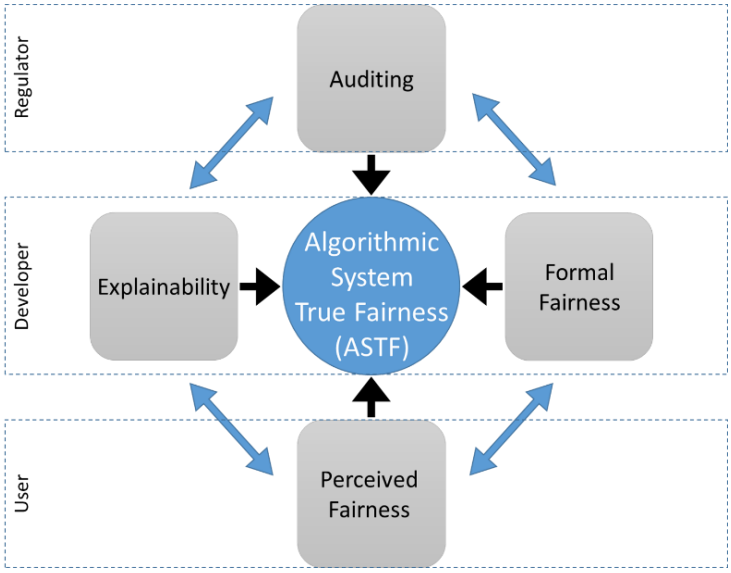
- Προκατάληψη του προγραμματιστή (developer bias): εκείνοι που δημιουργούν τα αλγοριθμικά συστήματα κάνουν πολλές επιλογές κατά τη διαδικασία, από την επιλογή των δεδομένων εκπαίδευσης έως την επιλογή του αλγόριθμου μάθησης που θα εφαρμοστεί στα δεδομένα και στον τρόπο με τον οποίο αξιολογείται το αλγοριθμικό μοντέλο για την απόδοσή του. Οι προγραμματιστές ενδέχεται να κακομεταχειριστούν κατά λάθος τα δεδομένα ή / και να μην ενδιαφερθούν για την άποψη των τελικών χρηστών του συστήματος. Με τον ίδιο τρόπο, οι δικές τους κοσμοθεωρίες και προκαταλήψεις (π.χ. κοινωνικά στερεότυπα που επηρεάζουν την κρίση τους) μπορεί να επηρεάσουν τις επιλογές τους.

- Προκατάληψη χρήστη (user bias): Τέλος, οι χρήστες δεν χρησιμοποιούν πάντοτε τα αλγοριθμικά συστήματα κατάλληλα. Ένα κοινό πρόβλημα είναι η "μεταφορά της μεροληψίας του περιεχομένου". Αυτό συμβαίνει όταν ένας χρήστης εφαρμόζει ένα αλγοριθμικό σύστημα σε ένα περιβάλλον που είναι διαφορετικό από την προβλεπόμενη χρήση του. Ένα παράδειγμα θα μπορούσε να είναι ένα σύστημα για την αξιολόγηση των αιτήσεων τραπεζικών δανείων, το οποίο σχεδιάστηκε και αναπτύχθηκε στο αμερικανικό πλαίσιο. Η εφαρμογή αυτού του συστήματος χωρίς συμμορφώσεις στο κοινωνικοοικονομικό πλαίσιο της Κύπρου θα μπορούσε να οδηγήσει σε απροσδόκητα προβλήματα.

  Τέλος, οι χρήστες πρέπει να γνωρίζουν ότι οι δικές τους συμπεριφορές μπορούν να επηρεάσουν τις συμπεριφορές ενός αλγοριθμικού συστήματος. Μέσα σε ένα σύστημα πρόσβασης πληροφοριών, όπως ένας μηχανισμός αναζήτησης ή ένα σύστημα συστάσεων, οι χρήστες συχνά δεν γνωρίζουν ότι οι συμπεριφορές τους παρακολουθούνται και χρησιμοποιούνται για την παροχή ανατροφοδότησης στο σύστημα. Αυτό σημαίνει ότι εάν οι χρήστες αποδέχονται ασυμβίβαστα τα αποτελέσματα που είναι ακατάλληλα ή άσχετα (δηλαδή επιλέγοντας χωρίς κριτική κάθε φορά που παρουσιάζονται τα κορυφαία αποτελέσματα), το σύστημα δεν θα μάθει να βελτιώνεται. Με αυτό τον τρόπο, οι διαδεδομένες προκαταλήψεις αναπτύσσονται. Τα κορυφαία αποτελέσματα τείνουν να παραμένουν δημοφιλή στα συστήματα αναζήτησης και συστάσεων, όχι επειδή είναι αναγκαστικά τα καλύτερα αποτελέσματα, αλλά επειδή οι χρήστες δεν καταβάλλουν προσπάθεια να εξετάσουν περαιτέρω τον κατάλογο των αποτελεσμάτων προκειμένου να ανακαλύψουν νέες ή εναλλακτικές προτάσεις/εξόδους (Ο).

2.5 Αντιμετώπιση των προκαταλήψεων στα αλγοριθμικά συστήματα: Προώθηση της δικαιοσύνης, της ευθύνης και της διαφάνειας (FAT)

Σε αυτή την ενότητα, συζητάμε τον τρόπο με τον οποίο διαφορετικές ομάδες ενδιαφερομένων, μπορούν να βοηθήσουν στην προώθηση αλγοριθμικών συστημάτων που να αντιμετωπίζουν τους ανθρώπους *δίκαια*, και τα οποία μπορούν να θεωρηθούν *υπεύθυνα* για τις αποφάσεις που παίρνουν, λόγω των *διάφανων* συμπεριφορών τους (δηλαδή μπορούν να εξηγηθούν/ερμηνευτούν σε ένα άτομο). Όπως απεικονίζεται στην Εικόνα 19, συμμετέχουν τουλάχιστον τρεις κατηγορίες ενδιαφερομένων σε αυτή τη διαδικασία και οι δραστηριότητές τους είναι όλες απαραίτητες για να διασφαλιστεί ότι τα συστήματα είναι πραγματικά δίκαια: ρυθμιστικές αρχές, προγραμματιστές και χρήστες.



Εικόνα 19: Διαδικασίες και ρόλοι εμπλεκόμενων φορέων στην προώθηση FAT αλγοριθμικών συστημάτων.

Ρυθμιστικές αρχές (και άλλοι ελεγκτές) (Regulators)

Με την έννοια ρυθμιστικές αρχές, εννοούμε ένα ουδέτερο τρίτο μέρος που δεν εμπλέκεται στην δημιουργία του αλγοριθμικού συστήματος, αλλά είναι υπεύθυνες για την αξιολόγηση των συμπεριφορών του. Εδώ αναφέρουμε στον έλεγχο (auditing), ο οποίος, σύμφωνα με το λεξικό Merriam-Webster, είναι «μια μεθοδολογική εξέταση και ανασκόπηση»[29] των συμπεριφορών ενός αλγοριθμικού συστήματος, με έμφαση στην ανίχνευση μεροληπτικής συμπεριφοράς. Οι ερευνητές αναπτύσσουν μεθόδους για τη διεξαγωγή διαφόρων τύπων αλγοριθμικών ελέγχων, ανάλογα με τα χαρακτηριστικά του συγκεκριμένου συστήματος.[30]

Οι 'ρυθμιστικές αρχές' συνήθως να συνδέονται με τις κυβερνήσεις, έτσι ώστε να είναι σε θέση να αναλάβουν δράση για να παραμείνουν υπεύθυνοι οι ιδιοκτήτες των συστημάτων. Ένα παράδειγμα στην Ευρωπαϊκή Ένωση είναι η επιβολή του κανονισμού γενικής προστασίας δεδομένων (GDPR). Σύμφωνα με τα άρθρα 51-59, τα κράτη μέλη πρέπει να διαθέτουν ανεξάρτητη και δημόσια αρχή για τον έλεγχο της συμμόρφωσης και την αντιμετώπιση της μη συμμόρφωσης (δηλ. Της Υπηρεσίας του Επιτρόπου Πληροφόρησης)[31]. Αρκετά άρθρα με GDPR (π.χ. άρθρο 22) θέτουν όρια στα συστήματα που λαμβάνουν αυτοματοποιημένες αποφάσεις με βάση τα προσωπικά δεδομένα των πολιτών[32]. Συστήματα αλγορίθμων ΙΑ, τα οποία χρησιμοποιούν αλγόριθμο προσδιορισμού προφίλ χρηστών φαίνεται να υπόκεινται σε τέτοια ρύθμιση. Ωστόσο περιμένουμε να δούμε πώς θα εφαρμοστούν οι κανονισμοί αυτοί.

[29] https://www.merriam-webster.com/dictionary/audit

[30] Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, *22*.

[31] https://www.wired.co.uk/article/what-is-gdpr-uk-eu-legislation-compliance-summary-fines-2018

[32] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/what-does-the-gdpr-say-about-automated-decision-making-and-profiling/

Στο εγγύς μέλλον, αναμένεται ότι οι φορείς της βιομηχανίας θα χρησιμοποιηθούν και αυτοί για τον έλεγχο και τη ρύθμιση των αλγοριθμικών συστημάτων. Για παράδειγμα, ο Σύνδεσμος Προτύπων IEEE έχει σήμερα ένα υπό εξέλιξη πρότυπο για τις 'Αλγοριθμικές Μεροληψίες'[33]. Οι οργανισμοί θα μπορούν να λαμβάνουν πιστοποιητικό, δείχνοντας ότι έχουν κάνει τη δέουσα επιμέλεια όσον αφορά την ελαχιστοποίηση της αλγοριθμικής προκατάληψης στα συστήματα που αναπτύσσουν.

Πρέπει επίσης να σημειωθεί ότι παρατηρούμε συχνά άλλα μέρη, όπως τα περιοδικά και τους ερευνητές, να διενεργούν ελέγχους. Σε αυτή την περίπτωση, ο στόχος δεν είναι η επιβολή νομικών ρυθμίσεων αλλά η αύξηση της ευαισθητοποίησης σχετικά με τις συμπεριφορές που εισάγουν διακρίσεις σε αλγοριθμικά συστήματα. Για παράδειγμα, η φυλετική προκατάληψη στο σύστημα COMPAS, που χρησιμοποιήθηκε στις Ηνωμένες Πολιτείες για να βοηθήσει τους δικαστές να καθορίσουν τις ποινές τους, δημοσιοποιήθηκε για πρώτη φορά από δημοσιογράφους[34] οδηγώντας σε διάφορες νομικές ενέργειες κατά της χρήσης του.

Προγραμματιστές (Developers)

Εκείνοι που αναπτύσσουν αλγοριθμικές διαδικασίες και συστήματα είναι συνήθως οι μόνοι που πρόσβαση τους στον κώδικα είναι δεδομένη . Ωστόσο, σε πολλές περιπτώσεις, η πολύπλοκη φύση των διαδικασιών προβλέπει στην ανάγκη διεξαγωγής ανίχνευσης διακρίσεων. Για το σκοπό αυτό, οι ερευνητές μηχανικής μάθησης και οι επαγγελματίες έχουν αναπτύξει διάφορες τεχνικές διαδικασίες και δοκιμές για να αξιολογήσουν τη δικαιοσύνη στους αλγορίθμους τους. Αυτές οι δοκιμές είναι επίσημες αξιολογήσεις δίκαιης συμπεριφοράς, με την έννοια ότι είναι καθορισμένες διαδικασίες - οι οποίες συχνά αποτελούν επίσημη μορφή αλγορίθμων - για να εκτιμήσουν το βαθμό στον οποίο η συμπεριφορά του συστήματος μπορεί να εμφανίσει διακρίσεις σε ορισμένα άτομα ή κοινωνικές ομάδες. Αν διαπιστωθεί ότι ένα αλγοριθμικό σύστημα είναι απαλλαγμένο από διακρίσεις, λέγεται ότι έχει υποβληθεί σε μια εσωτερική πιστοποίηση δικαιοσύνης.

Ένα άλλο ζήτημα που αντιμετωπίζουν οι προγραμματιστές είναι αυτό της εξήγησης. Αν αναπτύσσεται ένας αλγόριθμος έτσι ώστε οι συμπεριφορές του να μην μπορούν να εξηγηθούν ή να ερμηνευθούν από έναν άνθρωπο, δεν μπορεί να γίνει τίποτα για να εξασφαλιστεί η δικαιοσύνη του. Ως εκ τούτου, με αυτή την έννοια, η εξήγηση είναι ένα απαραίτητο μέσο για το επιθυμητό τέλος (δηλαδή, τη δικαιοσύνη). Όπως απεικονίζεται στην Εικόνα 19, αυτές οι διεργασίες αλληλοσυνδέονται για τη διασφάλιση της πραγματικής δικαιοσύνης σε ένα αλγοριθμικό σύστημα.

Χρήστες (Users)

Ο τρίτος φορέας είναι ο χρήστης ενός αλγοριθμικού συστήματος. Είναι σημαντικό να συνειδητοποιήσουμε ότι ακόμα και αν ένα αλγοριθμικό σύστημα έχει πιστοποιηθεί μέσω επίσημων διαδικασιών / δοκιμών ως δίκαιο, δεν σημαίνει πάντα ότι ο χρήστης συμφωνεί. Έτσι, υπάρχει ένα είδος άτυπης δικαιοσύνης, που αφορά την αντίληψη του χρήστη για το σύστημα και τις συμπεριφορές του απέναντι στους ανθρώπους.

Η άποψη του χρήστη σχετικά με τη δικαιοσύνη είναι σημαντική, καθώς σχετίζεται με το πόσο εμπιστεύεται ή όχι το σύστημα και τα αποτελέσματά του. Οι χρήστες είναι πολλοί και έχουν διαφορετικές αντιλήψεις για την αμεροληψία, που σε ορισμένες περιπτώσεις διαμορφώνονται από τις δικές τους πεποιθήσεις, τις κοινωνικό-πολιτισμικές ταυτότητες, τις εμπειρίες ζωής κ.λπ. Οι ερευνητές προσπαθούν επί του παρόντος

[33] https://standards.ieee.org/project/7003.html
[34] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

να καταλάβουν τι είδους εξηγήσεις και ποια είδη πιστοποιήσεις δικαιοσύνης μπορούν να βοηθήσουν στην ανάπτυξη εμπιστοσύνης από το χρήστη προς το σύστημα. Η Εικόνα 19 απεικονίζει αυτή την αλληλεξάρτηση.

Εκπαιδευτικοί (Educators)

Εκπαιδευτικοί όπως εσείς οι ίδιοι, αποτελούν μια συγκεκριμένη ομάδα χρηστών, καθώς οι δικές σας εμπειρίες με αλγοριθμικά συστήματα ΙΑ θα επηρεάσουν έμμεσα τις εμπειρίες των άλλων. Όπως θα περιγράψουμε σε αυτό τον Οδηγό, παίζετε έναν βασικό ρόλο στην ευαισθητοποίηση των μαθητών σας σχετικά με τις κοινωνικές και πολιτισμικές προκαταλήψεις που συχνά εμφανίζονται σε δημοφιλή συστήματα ΙΑ, ακόμα και εκείνες που μπορεί να χρησιμοποιείτε στην τάξη. Αυτός ο Οδηγός στοχεύει να σας βοηθήσει να ενσωματώσετε τα πλέον σύγχρονα ευρήματα από την έρευνα σχετικά με τη δικαιοσύνη στα αλγοριθμικά συστήματα, σε πρακτικές δραστηριότητες που μπορείτε να χρησιμοποιήσετε για την καλλιέργεια αλγοριθμικής παιδείας μεταξύ των μαθητών σας.

3. Λογική

Το New Oxford American Dictionary ορίζει τη λέξη ‘εγγραμματισμός (literacy)’ ως ‘την ικανότητα να διαβάζεις και να γράφεις’, αλλά και ‘την γνώση σε ένα συγκεκριμένο κλάδο’. Η Ευρωπαϊκή Επιτροπή, UNESCO και άλλοι οργανισμοί, υπογραμμίζουν ότι η παιδεία σήμερα είναι επίσης η ικανότητα χρήσης των ΠΕΤ και των κοινωνικών μέσων. Ωστόσο, για να κατανοήσουμε τον εγγραμματισμό του 21ου αιώνα, πρέπει να αναγνωρίσουμε ότι οι αλγόριθμοι και η Τεχνητή Νοημοσύνη (Α.Ι.) έχουν όλο και μεγαλύτερες συνέπειες στη ζωή μας, ιδιαίτερα την ελευθερία, την ιδιωτικότητα και την πρόσβασή μας στις ευκαιρίες. Ως εκ τούτου, είναι σημαντικό να αναπτυχθεί αλγοριθμική παιδεία, δηλαδή, να αυξηθεί η ευαισθητοποίηση σχετικά με το ρόλο που παίζουν οι αλγόριθμοι και να προωθηθεί η δημόσια καταγραφή των επιπτώσεών τους στη ζωή μας.

Το σκεπτικό πίσω από αυτόν τον Οδηγό είναι, λοιπόν, η ευαισθητοποίηση των εκπαιδευτικών σχετικά με την ανάγκη ανάπτυξης των δεξιοτήτων των μαθητών για την κατανόηση του τρόπου με τον οποίο οι αλγοριθμικές λειτουργούν στα συστήματα ΙΑ, όπως η αναζήτηση στο Google, και να διαμορφώσουν την άποψη τους για τις διάφορες πληροφορίες. Ο Οδηγός περιλαμβάνει δραστηριότητες (προγράμματα μαθημάτων) για την ευαισθητοποίηση των μαθητών σχετικά με τις κοινωνικές και πολιτικές συνέπειες των αλγοριθμικών προκαταλήψεων στο κυπριακό πλαίσιο, τις οποίες μπορείτε να χρησιμοποιήσετε στις δικές σας αίθουσες διδασκαλίας. Ο αυξανόμενος αριθμός ερευνητικών μελετών δείχνει ότι παρόλο που υπάρχει η αντίληψη ότι οι αλγόριθμοι και η τεχνητή νοημοσύνη είναι αντικειμενικοί και ουδέτεροι, οι περισσότεροι από αυτούς τους αλγόριθμους δεν είναι μόνο άγνωστοι στο κοινό αλλά και πολλοί από αυτούς έχουν αποδειχθεί προκατειλημμένοι.

Από πού προέρχεται αυτή η προκατάληψη;

> Είναι απλό. Αυτοί που προετοιμάζουν σύνολα δεδομένων (dataset) ή / και δημιουργούν αυτούς τους αλγορίθμους είναι άνθρωποι, οι οποίοι μπορεί να έχουν τις προκαταλήψεις τους και να μην το γνωρίζουν. Αυτές οι προκαταλήψεις μπορεί να είναι εναντίον ατόμων διαφορετικού χρώματος ή άλλων μειονοτήτων, γυναικών ή ατόμων με ειδικές ανάγκες. Ως εκ τούτου, είναι σημαντικό για τους εκπαιδευτικούς και τους μαθητές να γνωρίζουν τα διάφορα συστήματα ΙΑ που χρησιμοποιούνται στην τάξη και βασίζονται σε αλγορίθμους. Η κατανόηση του τρόπου με τον

οποίο λειτουργούν αυτοί οι αλγόριθμοι και η προσοχή στο πως τους αντιλαμβάνονται διαφορετικές ομάδες μαθητών είναι σημαντική για να λαμβάνονται σωστές αποφάσεις.

Οι γενικές παιδαγωγικές αρχές αυτής της προσπάθειας είναι οι εξής:

1. Ο δάσκαλος έχει το καθήκον να καθοδηγεί τους μαθητές και να διευκολύνει την εκμάθηση μέσω καλών παιδαγωγικών στρατηγικών, να εξηγεί έννοιες, να δίνει παραδείγματα, να ενθαρρύνει την συζήτηση, να υπογραμμίζει σημεία, να υποβάλλει ερωτήσεις, να δίνει ανατροφοδότηση και να ζητά από τους μαθητές να εκτελούν συγκεκριμένες εργασίες.
2. Ο δάσκαλος χρησιμοποιεί συγκεκριμένα κριτήρια για την αξιολόγηση των παιδαγωγικών δραστηριοτήτων που αναπτύσσει και υλοποιεί.
    - Σε ποιο βαθμό η δραστηριότητα ενσωματώνεται με την πραγματικότητα;
    - Σε ποιο βαθμό οι μαθητές έχουν ευκαιρίες να παρακολουθήσουν και να προβληματιστούν για τις δραστηριότητες που ασχολούνται;
    - Σε ποιο βαθμό η καλλιέργεια αλγοριθμικής παιδείας συμβάλλει στην επίτευξη συγκεκριμένων στόχων (βλ. σημείο 3);
    - Γίνονται οι μαθητές πιο εξειδικευμένοι στο να αναγνωρίσουν τη μεροληψία με τη χρήση αλγορίθμων και στην πρόταση λύσεων;
    - Σε ποιο βαθμό οι μαθητές περνούν χρόνο με άλλους μαθητές σε μια εργασία;
    - Σε ποιο βαθμό οι μαθητές βελτιώνουν την ικανότητά τους να διαπραγματεύονται λύσεις για αλγοριθμική διαφάνεια με τους συμμαθητές τους;
    - Σε ποιο βαθμό οι μαθητές ασχολούνται με τις ηθικές, κοινωνικές και πολιτικές επιπτώσεις της αλγοριθμικής προκατάληψης;
    - Βοηθούν τους μαθητές οι δραστηριότητες στις οποίες συμμετέχουν στη δημιουργία πολλαπλών πολύπλοκων λύσεων σε αλγοριθμική προκατάληψη που μπορούν να αναλυθούν και να αξιολογηθούν για την αποτελεσματικότητά τους;
3. Ο δάσκαλος καλλιεργεί αλγοριθμική παιδεία, δηλαδή αξίες και πρακτικές που θα προετοιμάσουν τους μαθητές για τη χρήση των κοινωνικών μέσων και του διαδικτύου για να αντιμετωπίσουν την αλγοριθμική προκατάληψη και τις ηθικές, κοινωνικές και πολιτικές συνέπειές της. Ο ρόλος του δασκάλου είναι να δημιουργήσει παιδαγωγικούς χώρους στους οποίους να εμπλέξει την αλγοριθμική προκατάληψη στην τάξη, ανοίγοντας δρόμους για συγκεκριμένες δραστηριότητες και εργασίες, για την αντιμετώπιση αλγοριθμικής προκατάληψης με τρόπους που προωθούν το δημόσιο καλό.

4. Προγράμματα μαθημάτων

4.1 Στόχοι

Τα παρακάτω προγράμματα μαθημάτων έχουν σχεδιαστεί για επίπεδο Δευτεροβάθμιας Εκπαίδευσης και έχουν ως στόχο την εισαγωγή μαθητών (ηλικίας 14-18) σε βασικές αλγοριθμικές διεργασίες προκειμένου να αυξήσουν τις γνώσεις τους για την αλγοριθμική προκατάληψη. Πιο συγκεκριμένα, τα προγράμματα μαθημάτων θα βοηθήσουν τους μαθητές:

1. Να προσδιορίζουν την ευρεία εφαρμογή αλγορίθμων στην καθημερινότητά τους και να κατανοήσουν πως οι καθημερινές τους δραστηριότητες εξαρτώνται από αλγοριθμικές διαδικασίες.

2. Να δείξουν πως η αλγοριθμική προκατάληψη μπορεί να επηρεάσει είτε τις επιλογές είτε τις αποφάσεις τους και να χρησιμοποιήσουν στρατηγικές για να κατανοήσουν πως λειτουργεί ο αλγοριθμικός χειρισμός.
3. Να επιχειρηματολογήσουν για την σημασία της αλγοριθμικής διαφάνειας και να συνδέσουν διάφορα παραδείγματα αλγοριθμικής μεροληψίας σε θέματα ιδιωτικότητας, οικονομικού κέρδους, και κοινωνικής ισότητας.

4.2 Δραστηριότητες και υλικά

4.2.1 Αλγόριθμοι στην καθημερινότητά μας

Οι αλγόριθμοι βρίσκονται πίσω από πολλές αποφάσεις που παίρνουμε στην καθημερινότητά μας: βασιζόμαστε σε αλγορίθμους για να αποφασίσουμε τι να αγοράσουμε, που να φάμε, με ποιον να είμαστε φίλοι ακόμα και με ποιον να βγούμε ραντεβού. Επομένως, η κατανόηση του τι είναι ένας αλγόριθμος και του τρόπου με τον οποίο λειτουργεί, σημαίνει ότι έχουμε μια καλύτερη εικόνα για το πώς εξαρτάται η καθημερινότητά μας από τους αλγορίθμους.

Οι στόχοι αυτής της ενότητας είναι:

1. Η παροχή βασικών ορισμών των αλγορίθμων
2. Η εξερεύνηση της διαδεδομένης χρήσης των αλγορίθμων στην καθημερινή ζωή μας μέσω συγκεκριμένων παραδειγμάτων κοινών ψηφιακών μέσων.
3. Η εξήγηση του τρόπου με τον οποίο χρησιμοποιούνται οι αλγόριθμοι για τη συλλογή δεδομένων από τους χρήστες, προκειμένου να βελτιωθούν τα αποτελέσματά τους.

Εργασία 1: Τι είναι ένας αλγόριθμος;

Ένας αλγόριθμος είναι ένας κατάλογος με κανόνες που πρέπει να ακολουθηθούν για την επίλυση ενός προβλήματος. Οι αλγόριθμοι καθορίζουν διαδοχικά βήματα που πρέπει να ληφθούν για να επιτευχθεί το επιθυμητό αποτέλεσμα.

Συζητήστε με τους μαθητές τον ορισμό του αλγορίθμου χρησιμοποιώντας παραδείγματα αλγορίθμων που χρησιμοποιούμε στην καθημερινή μας ζωή (π.χ. φτιάχνοντας ένα κέικ, ακολουθώντας ένα σύνολο οδηγιών για να φτάσουμε στο πάρκο, φορώντας ρούχα κάθε πρωί, ακολουθώντας μια λίστα με οδηγίες για να φτιάξουμε ένα τραπέζι).

Ωστόσο, τα ψηφιακά εργαλεία που χρησιμοποιούμε στην καθημερινή μας ζωή τροφοδοτούνται με πιο σύνθετους αλγορίθμους που μας επιτρέπουν να είμαστε πιο αποτελεσματικοί στη λήψη αποφάσεων. Αυτοί οι αλγόριθμοι μειώνουν την πολυπλοκότητα των πληροφοριών που έχουμε γύρω μας σε λίγες επιλογές που απευθύνονται σε εμάς και στα ενδιαφέροντά μας. Για παράδειγμα, όταν το Amazon μας προτείνει βιβλία με βάση αυτά που έχουμε αγοράσει, χρησιμοποιεί έναν αλγόριθμο για να υπολογίσει ποια άλλα βιβλία διαβάστηκαν από αυτούς που πραγματοποίησαν παρόμοιες αγορές με εμάς. Ακόμα ένα παράδειγμα είναι όταν το Facebook μας προτείνει φίλους, εξετάζοντας τους φίλους των φίλων μας για να μας υποδείξει ότι ίσως να θέλουμε να συνδεθούμε με αυτά τα άτομα.

Οι αλγόριθμοι χρησιμοποιούνται σε όλους τους τομείς της πληροφορικής. Δώστε στους μαθητές παραδείγματα αλγορίθμων και στη συνέχεια ζητήσετε τους να σκεφτούν άλλα παραδείγματα τέτοιων αλγορίθμων. Παραδείγματα αλγορίθμων είναι:

- Η μηχανή αναζήτησης Google που χρησιμοποιεί έναν αλγόριθμο για να βρει τις καλύτερες αντιστοιχίσεις για τις λέξεις κλειδιά της αναζήτησης. Αυτός ο αλγόριθμος αποφασίζει ποιες σελίδες εμφανίζονται πρώτα όταν κάνετε αναζήτηση.
- Ο ιστότοπος της Amazon χρησιμοποιεί αλγόριθμους για να καθορίσει τα αποτελέσματα εύρεσης και την τιμή των προϊόντων.

Εργασία 2: Πώς / γιατί οι αλγόριθμοι δημιουργούν φυσαλίδες;

Οι αλγόριθμοι ενισχύουν τη διαδικασία λήψης αποφάσεών, βοηθώντας μας να πετύχουμε αυτό που θέλουμε, γιατί συνεχώς τους λέμε ποιοι είμαστε και τι μας αρέσει. Οι αλγόριθμοι είναι αυτό που κάνει τις συσκευές μας ‘έξυπνες’. Κάνουν τις συσκευές μας να φαίνονται σαν μια μηχανή που σκέφτεται ότι σκεφτόμαστε εμείς. Στην πραγματικότητα, όμως, είναι ο αλγόριθμος που έχει ‘εκπαιδευτεί’ από τη συμπεριφορά μας και τις επιλογές μας που κάνουν τη μηχανή ικανή να γνωρίζει ή να προβλέψει τι θέλουμε να κάνουμε. Οι συνήθειες και οι προτιμήσεις μας είναι μέρος της διαδικασίας που κάνει τους αλγορίθμους να λειτουργούν και να γίνονται πιο αποτελεσματικοί.

Ωστόσο, εάν τροφοδοτούμε συνεχώς έναν αλγόριθμο μόνο με τις προσωπικές επιλογές μας τότε αυτό που παίρνουμε πίσω είναι κάτι που ενισχύει όλες αυτές τις προσωπικές προτιμήσεις. Αυτό δημιουργεί μια ‘φούσκα’ γύρω μας.

- Κάθε άτομο συμπληρώνει την προσωπική του φούσκα φίλτρων (filter bubble), τι πιστεύει μπορεί να είναι μέσα (και επομένως έξω) από τη φούσκα του. Για παράδειγμα, το άτομο Χ πιστεύει ότι τα αντικείμενα που αγόρασε πρόσφατα είναι μέσα στην φούσκα του.

Εργασία 3: Ποιους τύπους προσωπικών δεδομένων είμαστε πρόθυμοι να αποκαλύψουμε για να λάβουμε εξατομικευμένα αποτελέσματα;

Δεδομένης της παρουσίας αλγορίθμων στη ζωή μας, είναι σημαντικό να τεθεί το ερώτημα: Ποια κριτήρια χρησιμοποιούν οι αλγόριθμοι για να αποφασίσουν τα αποτελέσματα που είναι πιο ενδιαφέρον για εμάς;

- Χρησιμοποιήστε Κάρτες Δεδομένων για να διερευνήσετε πως παίρνονται οι αποφάσεις από τους αλγορίθμους και το αντίκτυπο που έχουν στην ζωές μας. Χρησιμοποιήστε τις κάρτες για να ξεκινήσετε μια συζήτηση σχετικά με τις διάφορες πληροφορίες που παρέχουμε ως είσοδο σε αυτούς τους αλγόριθμους. Συχνά παρέχουμε αυτά τα δεδομένα εις γνώση μας, αλλά πολλές φορές οι πληροφορίες αυτές αποκαλύπτονται μέσω των δραστηριοτήτων μας.
- Ζητήστε από τους μαθητές να σκεφτούν πώς λειτουργεί η ’εξατομικευμένη’ διαφήμιση (personalisation). Δώστε παραδείγματα εξατομικευμένης διαφήμισης την οποία ζητήθηκε (ζητώντας από την Amazon να σας συστήσει βιβλία) και εξατομικευμένης διαφήμισης που δεν ζητήθηκε (μια διαφήμιση για ένα ξενοδοχείο στις Βρυξέλλες που εμφανίζεται σε ιστότοπο ειδήσεων, αφού αναζητήσετε αεροπορικά εισιτήρια στις Βρυξέλλες).
- Συζητήστε με τους μαθητές σας τους διάφορους τύπους προσωπικών δεδομένων (π.χ. ηλικία, φύλο, θρησκεία, εισόδημα, εκπαίδευση, επάγγελμα), τι είδους δεδομένα αξίζουν περισσότερο για εμάς και για τις εταιρείες που χρησιμοποιούν τα δεδομένα αυτά. Ζητήστε από τους μαθητές να σκεφτούν τις ακόλουθες ερωτήσεις:
    - Ποιες πληροφορίες μοιράζονται όταν χρησιμοποιούν ορισμένες από αυτές τις μηχανές αναζήτησης / ψηφιακές εφαρμογές;
    - Τι είδους πληροφορίες θεωρείτε πιο σημαντικές;

- Ποιες πληροφορίες πιστεύετε ότι είναι πιο σημαντικές για τις εταιρείες;
- Τι γνωρίζετε για τον τρόπο συλλογής αυτών των δεδομένων από τις εταιρείες, και πώς χρησιμοποιούν αυτά τα δεδομένα;

Προτεινόμενα βίντεο για συζήτηση:

- Shoshana Zuboff για τον εποπτικό καπιταλισμό VPRO Ντοκιμαντέρ (Shoshana Zuboff on surveillance capitalism | VPRO Documentary)
- Η ελευθερία είναι ψέμα (Free is a Lie)
- Ζώντας με Αλγορίθμους. Γιατί θα πρέπει να νοιαζόμαστε για τους αλγόριθμους; (Living with Algorithms; Why should you care about algorithms?)

4.2.2 Πώς λειτουργούν οι αλγόριθμοι;

Οι αλγόριθμοι 'μαθαίνουν' και γίνονται πιο αποτελεσματικοί, σχετικά με τις ερωτήσεις που τους ρωτάμε. Όταν ρωτάμε κάποιες ερωτήσεις και στη συνέχεια δείχνουμε προτίμηση για ορισμένες απαντήσεις / αποτελέσματα, τότε ο αλγόριθμος 'μαθαίνει' ότι αυτό πρέπει να συνεχίσει να μας δίνει πίσω σαν αποτελέσματα. Δημιουργούμε την πραγματικότητα που μας προσφέρουν οι αλγόριθμοι. Οι αλγόριθμοι αντικατοπτρίζουν τον κόσμο που τους δείχνουμε. Όταν χρησιμοποιούμε αλγόριθμους για να βλέπουμε, να κατανοούμε και να παρουσιάζουμε ανθρώπους, τότε ο αλγόριθμος θα μετατρέψει αυτόματα τους ανθρώπους σε κατηγορίες, προφίλ και τύπους.

Με άλλα λόγια, όταν τα δεδομένα που χρησιμοποιούνται για την 'εκπαίδευση' του αλγορίθμου είναι προκατειλημμένα ή αντιπροσωπεύουν μια συγκεκριμένη άποψη μιας κοινωνίας ή μιας κοινότητας ή του κόσμου τότε το αποτέλεσμα είναι επίσης προκατειλημμένο. Αυτό το αποτέλεσμα χρησιμοποιείται στη συνέχεια για τη λήψη περαιτέρω αποφάσεων. Επομένως, δημιουργείται ένας κύκλος ανατροφοδότησης που είναι δύσκολο να σπάσει. Το ερώτημα που έχει προκύψει είναι πώς αυτός ο κύκλος ανατροφοδότησης επηρεάζει τη ζωή μας και το μέλλον μας, αν δεν το γνωρίζουμε;

Οι στόχοι αυτής της ενότητας είναι:

1. Να ερευνήσουμε τις διαφορετικές εκφράσεις αλγοριθμικής προκατάληψης.
2. Να διερευνήσουμε πώς η εισαγωγή των διαφόρων μεταβλητών αλλάζει τα αλγοριθμικά αποτελέσματα.
3. Να επιδείξουμε πώς η χρήση αλγορίθμων μπορεί να διαιωνίσει κοινωνικά προβλήματα όπως η προκατάληψη, η μισαλλοδοξία και οι διακρίσεις.

Εργασία 4: Επίδειξη της φούσκας φίλτρου

Τα αλγοριθμικά συστήματα χρησιμοποιούν προσωπικά δεδομένα για να φιλτράρουν ή/και να ταξινομούν αυτόματα το περιεχόμενο με βάση το προφίλ του χρήστη, για να καθοδηγούν τους χρήστες στο πιο σχετικό υλικό.

- Η μηχανή αναζήτησης χρησιμοποιεί δεδομένα όπως η τοποθεσία σας και η γλώσσα σας για την παροχή αποτελεσμάτων. Εάν είστε συνδεδεμένοι στη μηχανή αναζήτησης, τότε άλλες πληροφορίες, όπως το φύλο και η ηλικία σας, χρησιμοποιούνται ως τρόπος για τον προσδιορισμό των καλύτερων αποτελεσμάτων για εσάς.

- Επίσης, η μηχανή αναζήτησης χρησιμοποιεί άλλα δεδομένα που έχετε δώσει άθελά σας (γεωγραφική θέση και κίνηση, προηγούμενες αναζητήσεις, προτιμήσεις σας για ορισμένους ιστότοπους), προκειμένου να καθορίσει το τι θα θέλατε να δείτε στην κορυφή των αποτελεσμάτων αναζήτησης.

Η εκπαιδευτική επίδειξη (demo) 'Φυσαλίδας φίλτρου (Filter Bubble)' αναπτύχθηκε για να καταδείξει πως οι έξοδοι των αλγορίθμων θα μπορούσαν να διαφέρουν ανάλογα με την είσοδο. Το demo έχει δύο λειτουργίες την 'Explicit' και την 'Implicit'. Η λειτουργία 'Explicit' χρησιμοποιεί βασικές πληροφορίες που δίνει εις γνώση του ο χρήστη, όπως το φύλο και η ηλικία. Η λειτουργία 'Implicit' χρησιμοποιεί πληροφορίες που συλλέγονται από τη διατύπωση του ερωτήματος, όπως η πρώτη γλώσσα του χρήστη (υποδηλώνοντας τη γεωγραφική του θέση, το πολιτισμικό υπόβαθρο, κλπ.) ή τα επίθετα / επιρρήματα που χρησιμοποίησε ο χρήστης (υποδηλώνοντας τους στόχους ή τις συμπεριφορές του).

Οι μαθητές μπορούν να χρησιμοποιήσουν τη φούσκα φίλτρων και να διερευνήσουν πώς η αλλαγή της εισόδου τους επηρεάζει τα αποτελέσματα που παρέχει η μηχανή αναζήτησης.

Οι φοιτητές μπορούν επιλέγοντας ένα θέμα αναζήτησης από μία λίστα στην αρχική σελίδα του demo να αλληλεπιδράσουν με τα αποτελέσματα της αναζήτησης τους. Για παράδειγμα, αν το θέμα είναι η Κύπρος και το αρχικό ερώτημα είναι το όνομα στα αγγλικά ('Cyprus') ενώ οι παραλλαγές είναι σε διαφορετικές γλώσσες , Τουρκικά και Ρωσικά). Κάνοντας κλικ σε μια παραλλαγή τροποποιείται το μοντέλο χρήστη που χρησιμοποιείται για να φιλτράρει τα αποτελέσματα, ενημερώνοντας το κείμενο ή τις εικόνες που εμφανίζονται. Αφού ο μαθητής εξερευνήσει τουλάχιστον μία παραλλαγή, ενεργοποιείται το κουμπί για να μεταβεί στη σελίδα Επεξηγήσεων. Στη σελίδα Επεξήγηση οι μαθητές μπορούν να βρουν εξηγήσεις για το φαινόμενο της φούσκας φίλτρου, όπως: ένα σύντομο βίντεο που δείχνει τη διαδικασία φιλτραρίσματος και αναδιάταξης με εικόνες, ένα διαδραστικό τμήμα που μιμείται τα αποτελέσματα αναζήτησης (επιτρέποντας στο χρήστη να αλλάξει τα χαρακτηριστικά του προφίλ χρήστη και να δει τις αλλαγές αμέσως) και κείμενο με δυναμικές φράσεις που εξαρτώνται από το προφίλ του χρήστη και το θέμα. Υπάρχουν επίσης ορισμένες συμβουλές για την κατανόηση και τη διαχείριση των προσωπικών πληροφοριών που παρέχονται σε μια πλατφόρμα.

Εργασία 5: Συζήτηση

Οι μηχανές αναζήτησης και τα κοινωνικά δίκτυα φιλτράρουν πληροφορίες και δεδομένα σχετικά με εμάς που μας παρουσιάζουν με συγκεκριμένο τρόπο, ανάλογα με το ποιος πιστεύουν ότι είμαστε.

Αυτό είναι θετικό ή αρνητικό;

✔ Είμαστε εκτεθειμένοι σε πληροφορίες και επιλογές που βρίσκονται κοντά στα προηγούμενα ενδιαφέροντά μας (όπως αυτά έχουν καταχωρηθεί στο ιστορικό μας στο διαδίκτυο).

✘ Είμαστε περικυκλωμένοι σε μια προσωπική σφαίρα πληροφοριών, που ανακυκλώνονται τα ίδια πρόσωπα, πληροφορίες, νέα και ενδιαφέροντα.

✘ Οι εταιρίες και οι κυβερνήσεις έχουν στη διάθεσή τους λεπτομερείς πληροφορίες για εμάς => είμαστε υπό συνεχή επιτήρηση.

Συνέπειες της ζωής σε μια φούσκα φίλτρου:

- Δεν είμαστε εκτεθειμένοι σε πληροφορίες που η μηχανή / αλγόριθμος αναζήτησης υπολογίζει ως απόκλιση ή ως «αμήχανη» πληροφορία.
- Δεν δεχόμαστε νέα, διαφορετικά ερεθίσματα και πληροφορίες (προκατάληψη).

Ενθαρρύνετε τους μαθητές να σκεφτούν πώς θα δημιουργήσουν τη δική τους Φούσκα Φίλτρου προκειμένου να επιδείξουν αλγοριθμική προκατάληψη.

4.2.3 Αλγοριθμική Διαφάνεια

Οι αλγόριθμοι είναι ένα σύνολο βημάτων που οδηγούν σε ένα αποτέλεσμα. Για παράδειγμα ένας αλγόριθμος που χρησιμοποιούν οι άνθρωποι στην καθημερινή ζωή είναι μια συνταγή. Η αλγοριθμική διαφάνεια έχει πρόσβαση σε όλα τα βήματα του αλγορίθμου. Σκεφτείτε μια συνταγή, ένα παράδειγμα ενός πλήρως διαφανούς αλγορίθμου είναι ο μάγειρας που έχει πρόσβαση σε όλα τα απαραίτητα συστατικά και τα βήματα που θα οδηγήσουν στο επιθυμητό νόστιμο φαγητό.

Οι στόχοι αυτής της εργασίας είναι:

1. Να διερευνηθεί το ζήτημα της αλγοριθμική διαφάνειας και τι μπορεί να σημαίνει για τους καθημερινούς χρήστες.
2. Να συζητηθούν δεοντολογικά διλήμματα στη χρήση αλγορίθμων (προστασία της ιδιωτικής ζωής, επιδίωξη κέρδους, διακρίσεις)
3. Να αναπτυχθούν πρότυπα και οδηγίες για τη χρήση ψηφιακών εργαλείων που λειτουργούν με αλγόριθμους.

Εργασία 6: Οι απόψεις των συμμετεχόντων σχετικά με τη διαφάνεια του αλγορίθμου

Ζητήστε από τους μαθητές τις απόψεις τους για τη διαφάνεια στους αλγορίθμους.

- Είναι σημαντικό οι αλγόριθμοι να έχουν διαφάνεια; Αν ναι, γιατί;
- Πως πρέπει να είναι ένας διαφανής αλγόριθμος; Πώς πρέπει να γνωστοποιείται αυτό στους καθημερινούς χρήστες ψηφιακών εργαλείων;
- Συζητήστε τις προτάσεις που ενδέχεται να έχουν οι συμμετέχοντες σχετικά με τις αλλαγές που θα μπορούσαν να γίνουν στον τρόπο λειτουργίας του διαδικτύου.

Πιθανή εργασία με τις ακόλουθες πηγές από τον ιστότοπο της UNBIAS (π.χ., τι μπορούν να κάνουν οι γονείς; Βασικές ιδέες για το GDPR, αρχές για υπεύθυνους αλγορίθμους).

Περισσότερες πηγές για συζήτηση με θέμα την Αλγοριθμική Διαφάνεια:

- Αρχές για τους υπεύθυνους αλγόριθμους και σημείωση των κοινωνικών επιπτώσεων (Principles for Accountable Algorithms and a Social Impact Statement for Algorithms).
- Ασφάλεια των παιδιών στο διαδίκτυο: Ένας πρακτικός οδηγός για τους παροχείς κοινωνικών μέσων και διαδραστικών υπηρεσιών (Child Safety Online: A Practical Guide for Providers of Social Media and Interactive Services).
- Οδηγός για το GDPR (Guide to the GDPR).

Περαιτέρω πηγές για το διαδίκτυο και τις διακρίσεις:

- Η Google άναψε φωτιές με τα αποτελέσματα αναζήτησης ‘ρατσιστικών’ εικόνων για ‘μη εξειδικευμένα μαλλιά’ (Google under fire over 'racist' image search results for 'unprofessional hair').

- ‘Νομίζω ότι η μαυρίλα μου παρεμβαίνει’: Η αναγνώριση προσώπου δείχνει φυλετική προκατάληψη; (‘I think my blackness is interfering’: does facial recognition show racial bias?).
- Η HP διερευνά τις υποθέσεις των ‘ρατσιστικών’ υπολογιστών (HP Investigates Claims of 'Racist' Computers).

Πρακτικά εργαλεία για σωστή χρήση των κοινωνικών μέσων:

- Σε προγράμματα ανώνυμης περιήγησης
- Χρήση πολλών μηχανών αναζήτησης
- Χρήση διαφορετικών υπηρεσιών (όχι η ίδια εταιρεία, όχι απαραίτητα εμπορικές υπηρεσίες)
- Προβληματισμός για όταν εκτίθενται σε αυτοματοποιημένες προτάσεις ή διαφημίσεις: Ρωτάμε: ‘γιατί βλέπω αυτήν τη διαφήμιση;’

Ως ψηφιακοί πολίτες, απαιτούμε τα δικαιώματά μας:

- Ζητούμε διαφάνεια, έχουμε το δικαίωμα να γνωρίζουμε τι γνωρίζουν για εμάς
- Ζητούμε να γνωρίζουμε ποιος αποφασίζει και πώς λαμβάνονται οι αποφάσεις σχετικά με τον τρόπο με τον οποίο πρόκειται να χρησιμοποιηθούν τα δεδομένα μας
- Θέλουμε να έχουμε λόγο σε αυτές τις αποφάσεις

4.3 Αξιολόγηση

Οι στόχοι της Αξιολόγησης είναι να:

1. Παρακολουθεί την μάθησης των μαθητών (καθορίζει την ανάπτυξη, τα δυνατά και τα αδύνατα σημεία του κάθε μαθητή)
2. Δίνει ανατροφοδότηση στους εκπαιδευτικούς για να βελτιώσουν την διδασκαλία τους
3. Δίνει ανατροφοδότηση στους μαθητές για να βελτιώσουν την μάθηση τους.

Οι εκπαιδευτικοί μπορούν να αξιολογήσουν τις επιδόσεις των μαθητών χρησιμοποιώντας τόσο τη διαμορφωτική όσο και την αθροιστική αξιολόγηση. Κατά την εκτέλεση των πλάνων διδασκαλίας οι εκπαιδευτικοί μπορούν να διεξάγουν δραστηριότητες - είτε ατομικές, είτε ομαδικές - προκειμένου να αξιολογήσουν το επίπεδο κατανόησης των μαθητών και την ανάγκη τους για περαιτέρω διευκρινίσεις.

Η διαμορφωτική αξιολόγηση (formative assessment) είναι περισσότερο διαγνωστική παρά εκτιμητική. Πιο συγκεκριμένα, ο στόχος της διαμορφωτικής αξιολόγησης είναι να βοηθήσει τους μαθητές να εντοπίσουν τα δυνατά και τα αδύνατα σημεία τους και να στοχεύσουν στους τομείς που χρειάζονται δουλεία. Επίσης, η διαμορφωτική αξιολόγηση στοχεύει να βοηθήσει τους εκπαιδευτικούς να βελτιώσουν και να προσαρμόσουν τις μεθόδους διδασκαλίας τους αναγνωρίζοντας που δυσκολεύονται οι μαθητές τους.

Από την άλλη, ο στόχος της αθροιστικής αξιολόγησης (summative assessment) είναι να εκτιμηθεί η εκμάθηση των μαθητών και η ακαδημαϊκή επίδοση τους σε μια συγκεκριμένη χρονική στιγμή (στο τέλος του μαθήματος, στο τέλος της ακαδημαϊκής χρονιάς) συγκρίνοντας την με ορισμένα πρότυπα ή δείκτες αναφοράς. Επιπρόσθετα, η αθροιστική αξιολόγηση αποσκοπεί στην αναγνώριση κοινών ελλείψεων στην εκμάθηση των μαθητών και στην αναγνώριση των δυνατών και τον αδύνατων σημείων του προγράμματος μαθημάτων. Τέλος, βοηθάει τους εκπαιδευτικούς να καταλάβουν αν υπάρχει ανάγκη ανάπτυξης περαιτέρω δραστηριοτήτων ή / και η αλλαγή του τρόπου διδασκαλίας.

Παραδείγματα διαμορφωτικής αξιολόγησης είναι:

- Ζητήστε από τους μαθητές να δημιουργήσουν ένα οπτικό χάρτη με όσα έμαθαν
- Μικρά διαγωνίσματα μετά από κάθε μάθημα
- Συζήτηση στην τάξη μετά το τέλος κάθε μαθήματος
- Ασκήσεις για το σπίτι με δομημένη ανατροφοδότηση.

Παραδείγματα αθροιστικής αξιολόγησης είναι:

- ‘Τελική’ εξέταση μετά την ολοκλήρωση τους προγράμματος μαθημάτων
- Μία ομαδική εργασία. Δημιουργήστε ομάδες των 4-5 μαθητών και ζητήστε τους να παρουσιάσουν στην τάξη ένα συγκεκριμένο θέμα (μπορούν να επιλέξουν ένα σύστημα ΙΑ που χρησιμοποιούν καθημερινά και να προσπαθήσουν να καθορίσουν τις πληροφορίες που συλλέγει και να συζητήσουν αν θεωρούν το σύστημα δίκαιο). Στο τέλος κάθε μαθήματος παραχωρείστε λίγο χρόνο στους μαθητές σας για να εργαστούν στην ομαδική τους εργασία λαμβάνοντα υπόψη το μάθημα της μέρας.