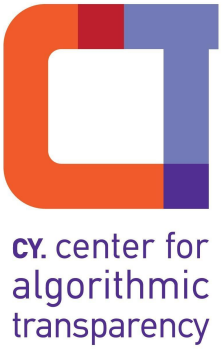

Document Title	**Summary of social and cultural parameters to guide bias detection work**
Project Title and acronym	Cyprus Center for Algorithmic Transparency (CyCAT)
H2020-WIDESPREAD-05-2017-Twinning	Grant Agreement number: 810105 — CyCAT
Deliverable No.	D3.3
Work package No.	WP3
Work package title	Understanding social and cultural consequences of algorithms
Authors (Name and Partner Institution)	Fausto Giunchiglia (UNITN) Jahna Otterbacher (OUC)
Contributors **(Name and Partner Institution)**	Veronika Bogin (UH) Alan Hartman (UH) Styliani Kleanthous (OUC) Tsvi Kuflik (UH) Avital Shulner Tal (UH)
Reviewers	Jo Bates (USFD) Kalia Orphanou (OUC) Lena Podoletz (UEDIN)
Status **(D: draft; RD: revised draft; F: final)**	F
File Name	D3.3_Summary_Bias_Detection_M12
Date	30 September 2019

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	4/9/2019	Jahna Otterbacher	jahna.otterbacher@ouc.ac.cy	Initial version
v2.0	16/9/2019	Jahna Otterbacher	jahna.otterbacher@ouc.ac.cy	Presented to partners at Annual Meeting
v3.0	27/9/2019	Jahna Otterbacher	jahna.otterbacher@ouc.ac.cy	Revised version

Abstract

Deliverable D3.3 provides a conceptual framework, using the concept of *diversity* as a lens to understand the biases of algorithmic systems. Using this diversity perspective, we describe a number of parameters upon which algorithmic system biases might manifest, as evidenced in the literature reviewed in D3.1 (literature review). Our framework is designed to guide system analysts and developers in bias detection work, with a particular emphasis on algorithmic information access systems used extensively in Cypriot society as well as internationally.

Keyword(s):	Algorithmic bias detection, diversity, perspective-taking, social and cultural bias

Contents

1. Executive Summary

D3.3 builds upon the work completed in the first two deliverables of WP3. In D3.1, we presented a comprehensive review of scholarly articles describing algorithmic system bias and the promotion of Fairness, Accountability and Transparency (FAT) across key application areas related to information access systems. This allowed us to understand the state-of-the-art in the field, both in terms of the problems detected by researchers, as well as the solutions proposed to address the issues examined. In contrast, D3.2 constitutes a dataset of "real world" case studies of algorithmic system biases reported in the popular press and social media. The current deliverable (D3.3) aims to develop a conceptual framework to aid in the detection of social and cultural biases in algorithmic information access systems.

After an extensive review of the literature (D3.1), we found that the issue of *diversity* - which is reflected in data, information, as well as the beliefs and behaviours of system users and developers - had largely been left out of the scientific discussion on algorithmic bias and FAT. In the context of this work, we take *diversity* to mean the co-existence of contradictory views and statements, some of which may be non-factual or referring to opposing beliefs or opinions (Giunchiglia et al., 2012). A discussion focused specifically on diversity and its relationship to algorithmic bias is long overdue, given that the most influential information access (IA) systems today have a global user base. Therefore, our conceptual framework for detecting and analyzing algorithmic biases is built on a "diversity perspective," which we argue is necessary for achieving a more holistic understanding of social and cultural biases in these systems, and thus, for generating meaningful solutions.

This document is structured as follows; in Section 2, we present a conceptual lens through which to study algorithmic system biases and the promotion of FAT in algorithmic systems. Specifically, we use the concept of *diversity* - in data, information and knowledge of the world - to develop a means to define and study algorithmic system bias. We characterize the problem space of algorithmic bias, through this diversity lens. In Section 3, we provide examples of how this lens can be used to analyze algorithmic bias in specific systems, using several case studies from the articles collected in our literature review (D3.1), which concern three types of algorithmic systems.

In Section 4, we present three "real world" case studies, emphasizing the CyCAT flagship systems/applications. These cases shall illustrate i) the types of biases known to occur in IA systems and their potential impact on Cypriot society, and ii) the use of our diversity framework in analyzing the problems of social and cultural biases. Finally, in Section 5, we summarize our findings.

2. Promoting FAT: A Diversity Perspective

Algorithmic systems are socio-technical in nature. For instance, machine learning and algorithmic systems are informed by human judgment at every step of the development process (Barocas et al., 2018). Training and evaluation datasets aim to capture aspects of the state-of-the world, and learning mechanisms are applied to create models based on them. Thus, systems reflect the biases of the societies in which they are developed. Some lead to discrimination, as in the much discussed case of the COMPAS system for predicting recidivism.[1] However, the goal of developing algorithmic systems is to introduce them into social contexts in hopes that they will be unbiased, acting fairly and achieving the required output in a just manner.

In today's world, we are faced with "hyper-diversity," as the Internet has long led us to a globalized environment. On any given day we may interact with systems and people, located across the world. This exposure to diversity has had a profound effect on information systems. When designing a system, we have an implicit model of its users, based on the development process we followed. Yet, when we deploy the system, we cannot anticipate who will use it and how. For practical reasons, we usually ignore the intentional or unintentional consequences that the system will have when introduced into a social context (Selbst et al., 2019). Thus, we may not consider that the actual users of our system will be diverse - and likely different - than the audience we initially had in mind.

A decade ago, the issue of diversity was much less critical, as the space and time limitations shielded us from it. However, this is no longer the case. All of us – and the systems we develop – are exposed to new diversity on a daily basis, manifested in language, data and knowledge. Arguably, the diversity of users of networked, intelligent algorithmic systems has exacerbated the issue of social and cultural bias. A system may show behaviours that deviate from what users expect, or what they consider to be normal with respect to their own context and perspective. When deviations in system behaviour are perceived, this then leads to discussions of whether the system is behaving in a manner that is fair. However, the challenge is that there is no single standard to which we can compare the behaviours of a given system; with a globalized user base, what is "normal" depends on many contextual factors, including one's socio-cultural environment and the prevailing values in a society (Dignum, 2017). As will be discussed, a lack of transparency means that a user or developer cannot appreciate the different perspectives of other stakeholders. Furthermore, the system will be used and interpreted within a social context, in which certain stakeholders have the power to define the "universal" norm, while others do not.

Transparency has long been recognized as a desirable property of an information system (Tyugu, 2016). Recently, researchers have emphasized its importance in ensuring that systems can be held accountable for harmful behaviours or the perception of unfairness (Datta et al., 2016; Diakopoulos, 2016). Yet research on algorithmic biases and FAT is dispersed across communities, with little consensus on definitions and approaches (see for example, the "21

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

Definitions of Fairness and their Politics" (Narayan, 2018) and “50 Years of Test (Un)fairness: Lessons for Machine Learning" (Hutchinson & Mitchell, 2019)).

Danks and London (2017), citing the need for a comprehensive conceptualization of algorithmic system bias, put forth a taxonomy, including biases of a computational origin, as well as those arising from inappropriate use of a system. They detailed five sources of algorithmic bias: i) training data, ii) algorithmic focus (i.e., differential usage of attributes in the training data), iii) processing (e.g., use of a statistically biased estimator in a model), iv) transfer context (i.e., application in a context differing from the one for which the system was developed) and v) interpretation bias (i.e., user misinterpretation of the system output).

We argue that the notion of diversity is key to understanding the above sources of bias. Since diversity is inevitable, perspective-taking (Galinsky & Moskowitz, 2000) (i.e., interpreting the world in someone else’s shoes), can be a tool for determining when bias is problematic and for whom. In this work, we motivate a “diversity perspective," taking into account the perspective-taking work from psychology (Davis, 1983), bringing it into the discussion on algorithmic system bias and transparency. At the same time, transparency is not sufficient to ensure the *fairness* of algorithmic systems. We need to remain cognizant that each person’s perspective is socially situated and that no one can see the “entire picture.” Drawing from Standpoint Theory (Stoetzler & Yuval-Davis, 2002), we can understand that some stakeholders have the power to declare their own perspectives as a universal norm, which can impede fairness.

In a globalized world, diversity drives the need for FAT and in particular, for *transparency*; without transparency one cannot detect or understand system biases. The current work provides a different view, and continues the discussion initiated by Hutchinson and Mitchell (2019) on Fairness in ML systems, with the aim of minimizing the gap between theory on FAT and its application in intelligent algorithmic systems.

2.1 Diversity and Awareness of Bias

Diversity is considered by many disciplines to be a positive attribute. Nelson (2014) stresses that diverse teams – in terms of skill set, education, work experiences, perspectives on a problem, cultural orientation, gender, etc. – produce better results compared to homogeneous teams. Similarly, Galinsky and colleagues (Galinsky et al., 2015) emphasize that the USA’s success as a nation is strongly based on the diversity of its immigrants. However, Nelson also stresses that stereotypes (e.g., based on race or gender) that we formulate in societies are potential threats to a diverse team’s success. This is due to the expectation of certain outputs and/or behaviours from particular social groups (e.g., male vs. female, African Americans vs. Caucasians) such that when a deviation occurs, unfair treatment can cause conflicts within the team.

Galinsky et al. (2015) advocate that transparency in work practices can help a diverse society to thrive and minimize conflicts that occur due to biases. As mentioned previously, the Internet today, along with its applications, such as social networks, search engines, recommender systems, decision support systems, etc., represents a diverse community that we cannot ignore. As

scientists and researchers of Web, social, and/or intelligent algorithmic systems, we are facing diversity in more than one form: the diversity in data on the Web, the diversity of humans involved in developing and using a system, diversity in the output/information delivered to the user, to name just a few factors.

The Fairness, Accountability and Transparency (FAT*) community acknowledges that promoting transparency is a direction towards minimizing the unwanted side-effects (such as stereotyping and biases) of diversity in data, humans and output, consequently promoting fairness. Baeza-Yates (2018), in a recent article in the Communications of the ACM on the different types of biases on the Web, concludes by stating that we can only reduce bias if we are aware that it exists. Thus, developers of algorithmic systems and data creators (e.g., crowdworkers), need to be conscious of the diversity – based on dimensions such as culture, race, gender, age, knowledge, etc. – of the potential consumers of their system's output and making their systems transparent at different levels. In other words, they need to first recognize that there are *alternative perspectives* beyond their own.

2.2 Perspective-taking: Promoting Transparency

Researchers in psychology have emphasized the importance of perspective-taking in occasions where significant diversity exists, (e.g., differences in the knowledge, political views, or cultural backgrounds of participants), aiming to reduce prejudice (Shih et al., 2009), minimize bias and stereotyping (Todd et al., 2011) and increase in- and out-group empathy (Galinsky & Moskowitz, 2000; Shih et al., 2009; Vescio et al., 2003). Perspective-taking studies require one person to view a situation or a behaviour from the other person's point of view (Galinsky & Moskowitz, 2000); hence, the relevance to diversity and bias research. Studies involving culturally diverse groups of people have illustrated the power of perspective-taking. In their studies, Vescio and colleagues (2003) found that individuals engaged in perspective-taking endorsed pro-African American attitudes, minimizing other individuals' bias towards African Americans in the group. Unlike humans, systems do not have the ability to change their perspective towards a situation, a statement or a belief. However, the humans involved in each step of the development of an algorithmic system (from requirement analysis to input training data and algorithm development) do have the ability, and need to consider perspective-taking as one approach towards promoting transparency in algorithmic systems.

In their work, Shih et al. (2009) examined participants' empathic feelings towards an Asian character, depicted in a scenario in which he tried to overcome a societal stereotype. Overall, the study involved Caucasian, African American, Hispanic and other non-Asian participants. It was found that participants who engaged in perspective-taking showed empathic feelings and liking for the character, thus illustrating that perspective-taking can reduce prejudice towards out-group members. Similarly, Galinsky and Moskowitz (2000), in three experimental studies, found that allowing participants to see a different perspective, decreased stereotypic biases toward out-group members, by enabling participants to associate themselves with the subject. The results showed that participants demonstrated both in-group as well as out-group favouritism. In summary,

perspective-taking research includes a number of studies consistent with the above findings, underscoring its promising potential application in the context of FAT* research.

Therefore, in this work, we adopt perspective-taking as a means to promote transparency. We propose a conceptual framework for discussing and understanding bias in algorithmic systems, considering variables such as the cultural, language, knowledge and life-experience diversity of the users, the developer and the observer/researcher. Within data and systems, such diversity results in the co-existence of competing and/or contradictory statements or views, some of which may be non-factual or referring to opposing beliefs or opinions (Giunchiglia et al., 2012). A system's user base may be global, serving individuals who perceive the world differently and do not interpret system behaviours the same way. Thus, algorithmic systems must take diversity into account to enhance user experience (Gu et al., 2017; Kunaver & Požrl, 2017). Given the challenges, our next steps are to relate the notions of diversity and bias, putting forth relevant definitions.

2.3 Definitions

2.3.1 Representing knowledge in the world

Diversity is manifested in the ways that implicit knowledge is represented, even when we limit the discussion to the "factual" aspects of knowledge (Jovchelovitch, 2002). The same entity (object/person/event) may be described in infinite ways across observers, varying by community, culture and language, or even life experience (Galinsky & Moskowitz, 2000). When we describe an entity, we choose the properties to use, which according to our background, will best characterize that entity. In other words, the resulting description can be a person's perspective towards a situation or an entity. These properties define the space, *S*, over which we shall eventually measure bias.

To illustrate the above consider the entity *snail*. As snails have been common in European and Mediterranean cuisine for thousands of years, individuals from such cultures would likely use the property "food" when describing snails. In contrast, a person from East Asia might relate snails to beauty products, rather than food. Still, the strength of association may vary by observer gender or age. This diversity of perspective is reflected in the data used to train systems. The text and multimedia shared via the Web, often used in training corpora, reflect our perspectives and experiences. Similarly, crowdsourcing often involves asking workers of various backgrounds to annotate or judge an entity relying on their own understanding of the world, thus embedding in the data one's perspective that might encompass stereotypes and biases.

2.3.2 Defining an unbiased point of reference

Another aspect of diversity that we need to consider, concerns the choice of the reference or standard: the unbiased point, *O*. According to the Oxford dictionary, a standard is "something used as a measure, norm, or model in comparative evaluations." However, the choice of the

reference is not common across all observers. Specifically, Jones and Nisbett noted that people can process information in different ways due to their divergent perspectives (Jones & Nisbett, 1971).

Continuing on the previous example, many individuals raised in the US or UK have a strong aversion to snails as food or otherwise.[2] Those who are Jewish or Muslim often share this aversion, as snails are neither kosher nor Halal. For such individuals, the unbiased area in the hyperspace representing the entity snail will allow for little variance with respect to dimensions such as “food" or “pet," as compared to individuals of other backgrounds.

2.3.3 Measuring bias

Diversity also has a relation to bias, in the choice of the metric, *M*, to express the deviation from a given point, to the reference. In contrast to the statistical tests discussed by Hutchinson and Mitchell (2019) for measuring bias, here we assume that an entity (i.e., incoming observation) and one’s reference, are represented as vectors in an *n*-dimensional space, where dimensions represent the properties used to describe the entity. One could measure the distance using any number of measures (e.g., Euclidean, Cosine or Manhattan distance). However, distance measures have properties that make them more or less informative given various considerations (e.g., dimensionality). This affects perception of the deviation and thus, whether the observation is deemed to be “biased." Table 1 summarizes the formal notation.

Definition	Notation
Space	S
Reference point	O
Metric: how the distance between a given point and O is quantified within S	M
Individual	i
Algorithmic Bias	AB
The system developer	D
The system observer	V
Context in which the facts of the world are represented	C
Individual i perceives the world with respect to her own context	C_i
S_i is the metric space in which i interprets the world, and O_i represents what is “normal" for i	$C_i =< S_i, O_i >$
A statement concerning the world	a
Statement a belongs to a specific context C	$a \in C$
The Bias Space, the context in which Bias is perceived and measured	$B =< C, M >$
Individual i observes a, from her own context, C_i. The bias of the statement a is given by its distance from the individual’s reference point O_i, with respect to her metric for measuring distance in S. The Perceived Bias of i is:	$PB_i(a) = \|\|a - O_i\|\|$
AB depends on the reference contexts of two parties	D, and V.
By default, the reference context of the system is that of its developer	$C_D =< S_D, O_D > \neq C_V$
Perceived Algorithmic Bias	$PAB_V(a) = \|\|a - O_V\|\| \geq 0$
Average System Bias	$Mean_{PAB_V} = \frac{1}{N}\sum_{k=1}^{N} PAB_V(a_k)$

Table 1: Formal representation.

[2] As evidenced by the popular saying that boys are “made of snails and puppydog tails," while girls are “made of sugar and spice and everything nice."

2.3.4 Defining and measuring algorithmic bias

We have seen that diversity relates to a general notion of bias in terms of: i) how facts of the world are represented in data and information; ii) the standard against which any incoming observation will be compared; and iii) how the deviation between the observation and the standard is measured. With the above in mind, we provide the following definitions. (See Table 1 for the formal notation of the below definitions.)

Reference Context. Before moving onto formally defining algorithmic bias, we need to acknowledge that each individual observes the world differently or from a different perspective. Hence, context is an important part of understanding diversity and bias. As a result, the contextual reference point (i.e., individual perspective) for what is "normal" differs from person to person. For example, take C to be the context in which the facts of the world are represented. Individual i perceives the world with respect to her own context, C_i , where
$C_i =< S_i , O_i >$. S_i is the metric space in which i interprets/describes the world, and O_i represents what is "normal" for her. From here on *Context* will be used to represent a person's individual perspective.

Bias. Whether characterizing a situation, an event or an object, we are making a statement about the state of the world. Bias is a property of a given statement. However, this bias is independent of whether the statement is true or false, and depends strongly on the context that intermingles the individual's point of view, with her means of making comparisons (i.e., distance metric).

Assume a statement a concerning the world, where $a \in C$. The Bias Space, the context in which Bias is perceived and measured, is $B = < C, M >$. When we observe a situation, an event or an object in the world we do so from our own context. Thus, the deviation of a statement (measured distance), used for describing what we experience at a given point in S, from the reference point O_i we have, can be considered as the bias of the statement. Therefore, the Perceived Bias of i is $PB_i (a) = \| a - O_i \|$.

Algorithmic Bias. Systems, like people, also make statements describing the state of the world. Algorithmic bias (AB) is the bias generated by an algorithmic system. Extrapolating from the above definition of bias, algorithmic bias depends on the reference contexts of two parties: the developer D, and the system observer V.[3]

By default, the reference context of the system is that of its developer. Thus, we often have that the context of the developer will be different from the context of the system observer:
$C_D = < S_D, O_D > \neq C_V$. This is a key cause of algorithmic bias: the system is built under the developer's context C_D, but its behaviours are interpreted under a different reference context, C_V .

For example, when a developer is adding specific rules in an interactive dating system (e.g., how recommendations are made to a user), the developer is acting according to her own reference

[3] The observer may or may not be a user of the system. We shall return to this point.

context (C_D). When a researcher (i.e., a system observer) is auditing the system's output given a specific input, then the researcher will interpret this output based on his own context (C_V).

Measuring Algorithmic Bias. We can measure the algorithmic bias of a given system by taking into account the deviation of the system-generated statement, a, from the observer's context or viewpoint, C_V , and thus, bias space, B_V . In this case, the Perceived Algorithmic Bias will be $PAB_V\ (a) = \| a - O_V \| \geq 0$. It is important to note that from the developer's perspective, the algorithmic bias of the system generated statement will be zero ($PAB_D(a) = \| a - O_D \| = 0$) assuming that she acts in good faith.

Continuing on the dating system example above, if we assume that the developer did not inject any intentional bias within the system rules (acting in good faith), the perceived bias of a system output $PAB_D(a)$ will be zero. However, when an observer is interacting with the system, his perceived bias for a system output $PAB_V(a)$ will be greater than zero, given that he has a diverse background, as compared to the developer.

Measuring Average System Bias. Finally, it should be noted that algorithmic bias can also be measured across a representative sample of N statements, $a_1, a_2,...a_N$, generated by the system. For instance, the mean system bias might be calculated, again from the perspective of the observer, as $Mean_{PAB_V} = \frac{1}{N} \sum_{k=1}^{N} PAB_V(a_k)$. As will be shown in Section 3, in the scientific literature on FAT, researchers are generally interested in characterizing the average bias of a given algorithmic system, rather than whether or not an individual algorithmic statement is biased.

2.4 Diversity, Value Judgments and Fairness

We have seen that algorithmic systems make *value judgments* concerning the world, which may or may not align with those of an observer of the system. An algorithm is an artefact produced by a human developer, on the basis of some reference context (C_D). As mentioned, while the developer has an implicit model of the user during the development process, in reality, users will be diverse and perhaps unexpected. Any observer of the system's behaviour brings in a second reference context (C_V), which typically differs from that of the developer, and it is under this reference context and bias space (B_V), where the evaluation of the system behaviours takes place.

It must be noted that the determination of the reference context is also a value judgment, because it defines not only how we perceive the world (i.e., the space in which we characterize what we see, S) but also what we perceive as being expected, normal, or unbiased (O). One's reference context - and thus, bias space (B) - is a result of her culture and upbringing, and may change (at least a bit) over time, with life experience. Therefore, it is clear that the diversity of the world, and in particular, the differences between the reference contexts of people, is the root of algorithmic bias. Thus far, we have used perspective-taking as a means to *promote transparency* between the various stakeholders of an algorithmic system, and we have seen that each party - including the

system itself - makes value judgments concerning the world. Having emphasized the *Transparency* in FAT, we now briefly explain how this relates to *Fairness.*

2.4.1 Standpoint Theory: From transparency to the detection of (un)fairness

While it is inevitable that there will be some divergence between any two Contexts, since no two people are completely alike, it follows that there will be bias in data and systems as well. In fact, machine learning methods make use of patterns in data to identify useful, "benign" *biases* that correlate with real world phenomena. It is only when the system finds and uses (or otherwise outputs) *harmful biases* that the "algorithmic bias" is acknowledged and discussed outside of the development context. The distinction between which biases are harmful and which are benign is an issue of *fairness.* Transparency in the process of making this distinction is key to detecting *un*fairness.

While psychology introduced the concept of *perspectives*, sociologists often speak of *standpoints*. In essence, both of these terms refer to a "point of view" - a location from which a person observes and interprets the world. However, the assumptions grounding the concept of a *standpoint* distinguish it from the concept of *perspective*. Having emerged from Marxist traditions and making its way into Feminist Theory, Standpoint Theory is rooted in the explicit "assumption that all knowledge is socially situated" (Kvasny, 2006). Since each person's knowledge is created through social contexts, they can never see the whole picture, having access only to a situated, partial view of reality. In other words, the theory holds that there is no objective, universal *O.* That said, certain people and groups in society are positioned as to hold power over others, in the sense that they have the ability to constrain the choices available to others (Allen, A.,1998), by declaring their own O_V as being the societal norm.

While Standpoint Theory posits that it is not possible to objectively evaluate other standpoints, "legitimate knowledge" is often defined from standpoints of privileged groups such as white, straight, cisgender, wealthy men (Allen, B., 1998). Furthermore, this may not always be a conscious decision; Calvert and Ramsey (1996) argue that people often do not recognize their own privilege and social power. As a result, they may be unable to see how the effects stemming from their own actions and decisions may harm historically underserved groups (Kvasny, 2006).

It has been acknowledged that the tech sector is facing significant challenges in terms of maintaining a diverse workforce.[4] At the same time, there are ongoing reports of how the technology created and deployed tends to disproportionately harm people from historically marginalized groups (see for example, reports of racial bias in the Google search engine,[5] [6] or the iPhone X's failure to distinguish Asian faces in its Face ID algorithm[7]). Such problems may stem from the industry's culture; dominant groups have the power to influence or define the success of

[4] https://www.revealnews.org/article/heres-the-clearest-picture-of-silicon-valleys-diversity-yet/
[5]https://www.theguardian.com/technology/2016/apr/08/does-google-unprofessional-hair-results-prove-algorithms-racist-
[6]https://www.theguardian.com/technology/2016/jun/09/three-black-teenagers-anger-as-google-image-search-shows-police-mugshots
[7] https://www.mirror.co.uk/tech/apple-accused-racism-after-face-11735152

an organization (Luke, 1974), and their definition may naturally represent what is desirable from their standpoint, but fail to consider it from the standpoints of other groups.

Coming back to the behaviours of algorithmic systems, according to Standpoint Theory, even physical entities in the world can be represented in data and information in various manners depending on human interest. This is particularly true in terms of the features that one considers significant to analyze / report, and the level of detail used to describe each feature (Giere, 1999). Thus, a developer determines which aspects of the world to model, how to formalize the measurements, and whether or not the outputs of the system's model are accurate and acceptable. There is however, "no such thing as a complete [model] of everything" (Giere, 1999). In other words, there can be no system that simultaneously displays all stakeholder standpoints; thus, we must always consider whether the standpoints (i.e., value judgments) taken by a given system are potentially harmful to some people, and how.

3. Scientific Research on Algorithmic Bias

Having explored the relationship between diversity and bias, as well as the potential for perspective-taking and standpoint analysis to promote transparency and fairness in algorithmic systems, we now focus on the problem space, as described in the scientific literature. In particular, we examine how diversity and bias manifest and are measured within an algorithmic system. To this end, we review examples of FAT research across three domains, using the definitions of Section 2.3.

First, we provide a general characterization of algorithmic systems and their macro components, as well as of the role of the researcher as the system observer. A basic architecture is provided in Figure 1. First, the system receives input (I) for an instance of its operation; its operational component (i.e., algorithmic model - (M)) performs some computation based on the inputs provided and produces an output (O). The model learns from a set of observations of data (D) from the problem domain. It may receive constraints from third party actors (T), and/or fairness criteria (F), which modify the operation of the algorithmic model (M).

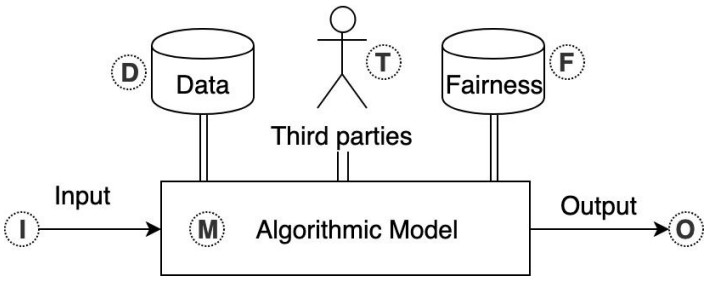

Figure 1: General architecture of an algorithmic system.

Recall that Algorithmic Bias (AB) depends on the reference contexts of the developer and the observer and this reference context will define their perspective towards the system's output. While the context of the developer C_D may be unknown, the context of the observer C_V is known

or implied by the manner in which the study is conducted, and serves as the context for the evaluation. Thus, FAT research often characterizes the average algorithmic bias AB of a system, as perceived by the researcher/observer, $PAB_V(a) = \| a - O_V \|$. With the above in mind, we analyze examples of the problem space, as presented in publications from three domains - text classification, search engines and recommender systems, leaving the discussion of the solution space for future work. When considering the problem space of each example we:

- Characterize the algorithmic system addressed, in relation to Figure 1.
- Identify the relevant dimension(s) of diversity.
- Detail the manner in which PAB_V is calculated.

3.1 Text classification

Problem. Given an input text (I), the goal is to assign one or more appropriate classes (O) that describe some attribute of the text. In the case of a classifier based on supervised learning (M), such as those in the examples below, the necessary data for training the model is a corpus of text from the domain of interest, in which each observation/text has been labelled with the correct class(es) (D).

Example 1 [TC1]. Dixon and colleagues (2018) trained a binary classifier (M) to label Wikipedia comments as being toxic/not toxic (O), based on the words in the textual comment (I). The *diversity dimension* of interest was minority status, including groups based on sexual orientation and religious affiliation, which were flagged by the use of sensitive identity terms (e.g., gay, black, atheist, Muslim) that appeared in the training corpus (D).

The concern was that sensitive words were associated more frequently with examples of “toxic" rather than “not toxic" comments, resulting in unintended biases in the classifier. Therefore, the authors proposed balancing the toxic/not toxic examples in the dataset (F). Average perceived algorithmic bias (PAB_V) was calculated based on error rate metrics (i.e., comparing the rates of false positives and negatives) when classifying comments containing sensitive words versus those not containing such terms. Thus, the reference point of the observer(O_V) was the classification error rate for texts containing no sensitive terms.

Example 2 [TC2]. Shen and colleagues presented a sentiment analysis scenario (Shen et al., 2018). Social media texts (I) were labeled as having positive/neutral/ negative affect (O). Various black-box algorithms (M) were trained on corpora from social media (D). The *diversity dimension* of interest was race. Specifically, the authors were concerned about *stylistic bias*, such that texts containing linguistic markers of African-American English (AAE), were often mislabeled as negative. They proposed to “neutralize" incoming texts, such that the algorithms would analyze them like a text of comparable context, but without sensitive terms (F).

A regression model was used to calculate average perceived algorithmic bias (PAB_V). Specifically, each sentiment algorithm was used to score the original datasets, as well as datasets in which sensitive words were neutralized. The regression model related the two sets of sentiment

scores, such that average PAB_V represented the average change in sentiment scores (i.e., change in regression coefficient). In other words, O_V referred to the sentiment scores of texts in which markers of AAE were absent.

3.2 Search engines

Problem. Given an input query from a user (I), a search engine returns a ranked list of documents (O), meant to be relevant to the user's information need. The algorithmic models (M) behind modern proprietary search engines such as Google or Bing are difficult to characterize, not only because they are protected trade secrets, but also because of their complexity. The data used to train the models (D) likely consists of a combination of curated relevance datasets, datasets collected from "the wild," as well as user history and profile, etc.

Example 1-2 [SE1,SE2]. Two recent studies addressed the diversity dimension of *gender*, in proprietary image search engines, considering search results (O) returned in response to queries (I) concerning the professions (Kay et al., 2015) as well as character traits (Otterbacher et al., 2017). Fairness constraints (F) were not suggested; rather, the aim was to document bias to raise user awareness.

In (Kay et al., 2015), the reference point of the observer, O_V, was derived from US labour statistics on a given profession (i.e., the gender distribution of workers in the profession). The authors noted that the choice of O_V was not neutral, but rather, reflected the biases of the offline world. This was used as a benchmark of gender bias in the search engine results (presented to US-based users). Average PAB_V was computed by comparing the online versus offline gender distribution in retrieved images, across a set of profession-related queries.

In contrast, to compare the gender distribution of images retrieved for a given character trait query (e.g., "sensitive person"), no offline reference point was cited in (Otterbacher et al., 2017). Instead, the authors compared the images retrieved on a given query, across four search engine markets (US, UK, South Africa and India). Average pairwise deviation was examined between the four markets, by comparing the gender distributions in images retrieved across a large set of queries demonstrating different observer's reference point O_V.

Example 3 [SE3]. Mowshowitz and Kawaguchi demonstrated a method for measuring search engine bias in a proprietary search engine (Mowshowitz & Kawaguchi, 2005). Thus, a fairness constraint (F) was not imposed on the system. For a large set of user-generated search queries (I), they created a "fair results set," consisting of results retrieved for a given query, across a number of alternative search engines. The diversity dimension of interest is *information diversity*. Put another way, O_V expects adequate diversity in search results, regardless of the topic of the query. Thus, average PAB_V was calculated, based on the deviation of the search results given by the engine being evaluated (O), from the search results of the "fair results set."

3.3 Recommender systems

Problem. Given the profile of a user and an expressed need (i.e., query) (I), the system generates a ranked list of recommendations (O), deemed to be most compatible with the user's interests/needs. The system's model (M) is trained on datasets (D) capturing all users' interactions with the system. Additional constraints may be added by third parties (T) who interact with / take decisions in the system.

Example 1 [RS1]. A recent study examined gender and racial biases in two freelance marketplaces, TaskRabbit and Fiverr (Hannák et al., 2017). In both systems, users receive ranked lists of candidates (O) in response to search terms concerning small jobs (I). Third parties leave textual comments as well as ratings for candidates (T), who have completed previous jobs. The assumption was that candidates with similar qualifications and history of experience, should receive similar rankings, regardless of their gender/race.

The authors considered the feedback of candidates from other users, the language of the textual reviews on candidates' work (i.e., the use of positive/negative adjectives), and the candidates' positions in resulting rankings for a given job. Regression models were used to study the relationship of gender/race with these variables. The average system bias was measured in terms of the coefficients on gender/race (statistical significance, effect size). Ideas were put forward, such as re-ranking candidates to ensure fair treatment (F).

Example 2 [RS2]. Another example (Eslami et al., 2017), presented a cross-platform audit of hotel recommendations. The systems studied return a ranked list of recommendations (O), in response to a user's profile and search terms (I), taking into consideration others' ratings/ reviews (T). *Credibility* was the dimension of interest, as the primary concern was that the system Booking.com skewed users' ratings of hotels, as compared to other systems. The authors compared customer ratings of over 1.500 hotels, by taking the average difference in ratings between any two platforms, and testing for statistical significance. Specific fairness constraints (F) were not mentioned; however, the authors noted that users often raise awareness of biases through textual reviews of the hotels.

3.4 Summary of observations

By design, all algorithmic models are set to express intentional bias (e.g., in a search engine, a systematic preference for "relevant" content over that which is deemed less relevant). However, some algorithmic biases are unintended and potentially problematic, such as those examined above. Table 2 summarizes the problem spaces examined in these cases. As can be seen, the diversity dimensions examined reflect the potential for algorithmic biases to result in harm, such as discrimination against particular social groups (based on characteristics such as race, gender, religion or sexual orientation), or providing information to users that is not balanced or credible.

Another observation from Table 2 is that, while the representation of knowledge (*S*) used in a given system, as well as the distance metric used (*M*), are dependent on the domain and the particular problem at hand, it is interesting to note a commonality in the choice of the reference point (O_V). In particular, all reflect an attempt to find a "cultural consensus" on the baseline, either through the use of crowd wisdom obtained through open Web and social media (e.g., TC1, TC2, SE3), comparable data within the same or another system (RS1, RS2) or official government statistics (SE1). In all the above examples, we can appreciate the importance of perspective-taking as a way to acknowledge the *diversity*, that exists in the data, as well as in the humans involved in system development, auditing and use. Although we have by no means provided an exhaustive list of examples and cases where diversity and biases exist in intelligent algorithmic systems, we have demonstrated that perspective-taking can be an approach to promoting transparency amongst the various stakeholders of such systems.

Diversity dimension	**Representation (*S*)**	**Reference point (O_V)**	**Distance metric (*M*)**
TC1: Minority status	Word vector	Classification performance on text without sensitive words	Error rate equality difference
TC2: Race	Word vector	Classification performance on text without sensitive words	Change in sentiment scores
SE1: Gender	Gender distribution in images retrieved	Gender distribution per labour statistics	Distance between online/offline gender distributions
SE2: Gender	Gender distribution in images retrieved	None assumed	Pairwise differences between search markets
SE3: Information	Distribution of URLs retrieved	"Fair results" set from multiple engines	One minus the similarity between fair/empirical results
RS1: Gender, race	Feature vector (text, rating, rank)	Ranking of worker with similar history in system	Coefficient on diversity attributes in model to infer rank
RS2: Credibility	Rating, source	Rating at other platforms	Differences across platforms

Table 2: Comparison of problem spaces in FAT case studies across domains.

4. Analysis of Case Studies of IA systems

Having articulated a diversity framework and illustrated how it applies to the cases of algorithmic system bias in the scientific literature, we now present an analysis of information access (IA) systems used in "everyday contexts." In particular, the case studies presented in this section represent some of the most common IA systems used by Cypriot end users.

In particular, this section examines the diversity of outputs from image and video search engines (3.1), search engine "autocomplete"/suggestion algorithms (3.2), and video recommendation systems on social media (3.3). In these case studies, we find sources of diversity in both the input and output. For instance, in a user's input query for an information artefact of interest, diversity sources include the different languages used to formulate a query, as well as the choice of phrasing. The output (i.e., search, autocomplete or recommendation results), may differ in the identities of the people depicted in images and videos, the ideological/political standpoint of the information presented, and various other characteristics (i.e., diversity dimensions) depending on the platform, content types, and keywords.

The examples below describe interactions with proprietary IA systems, which have been executed from the same Cypriot IP address, within the same week (10 - 17 September 2019).

4.1 Google image search engine

To examine issues of diversity and social/cultural bias in Web search, we simulated the search behavior of different people, namely those who are native speakers of Greek (i.e., primarily Greek-speaking Cypriots and Greek nationals) and Turkish (i.e., Turkish-speaking Cypriots and Turkish nationals), as well as those who use the Internet in English. The last group includes those who are native speakers of English (potentially nationals of an anglophone country) as well as those who are not native speakers but choose to conduct searches in English, as English has often been considered the lingua franca of the Internet. In other words, in the terms of the framework presented in Section 2, we aim to observe the system's behaviour, from the point of view of three reference contexts of simulated users: Greek-speaking Context (C_G), Turkish-speaking Context (C_T), and English-speaking Context (C_E).

To examine the way information about Cyprus is presented to those using search engines, we conducted a series of image searches for Cyprus on the Google Images engine. We conducted the search in English ("Cyprus"), Greek ("Κύπρος"), and Turkish ("Kıbrıs"), taking screenshots of the first 20 results that appear in the search engine results page (SERP), and compare our findings across the different languages. Figure 2 displays the three SERPs. We do not know the developer's reference context (C_D) or unbiased reference point for each search (O_D). Therefore, we shall compare what each of the three simulated users observe, for the "Cyprus" query.

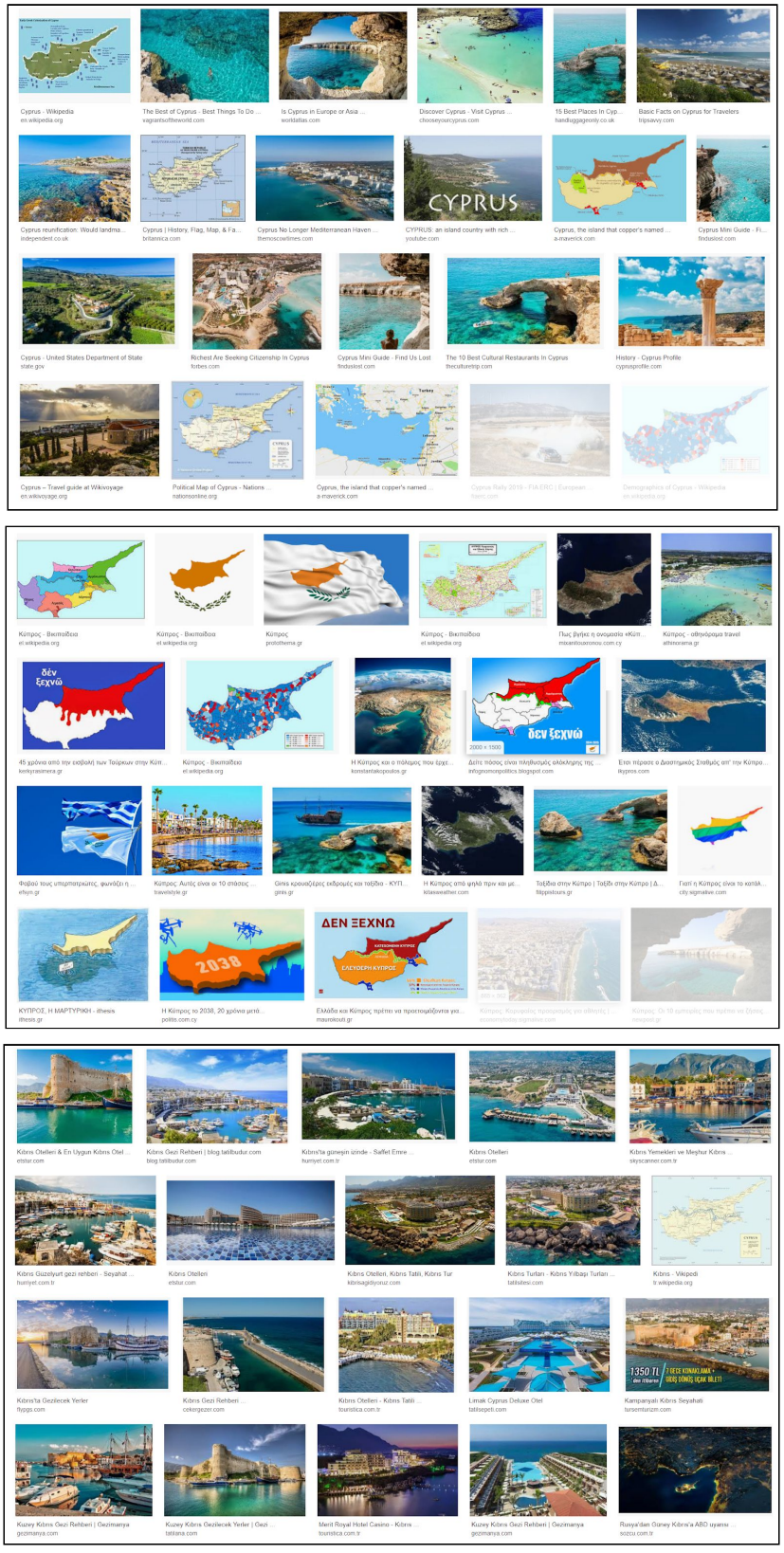

Figure 2: From top to bottom, the first 20 results from Google Image Search for the keywords: “Cyprus,” “Κύπρος,” and “Kıbrıs.”

4.1.1 Observations

Table 3 presents an analysis of the salient aspects in the images retrieved. It should be noted that these aspects do not constitute an exhaustive description of the images' content; rather, they highlight the social and cultural aspects that are sensitive within the Cyprus context. In addition, the aspects examined are not mutually exclusive. However, even with a cursory look at the overview of results (Table 3 and Figure 2), one observes that the images returned for the same search term are significantly different when the term is in different languages.

Aspect represented	**English**	**Greek**	**Turkish**
Coast/Beach/Sea	13	4	15
Map	5	6	1
Hotel/Resort	1	0	8
Historical site	2	0	6
Satellite image	0	4	1
Flag	0	3	0
Other	1	3	0
Descriptions of other aspects	Vineyard/hill (1)	Political cartoon with southern half of the island sunken (1), Island outline overlaid with LGBTQ+ flag (1), Illustration of island and drones (1)	-

Table 3: Aspects represented in the first 20 images returned, along with their frequencies, for queries in English ("Cyprus"), Greek ("Κύπρος"), and Turkish ("Kıbrıs").

The results often vary in that the images for each search represent different aspects of Cyprus, such as its beaches or its map (i.e., comparing across the columns in Table 3). However, it is also clear that certain aspects are emphasized for searches conducted in one language but not others (i.e., comparing across the rows in Table 3). For example, we observe many more images displaying the beach or the coast when the search is conducted in English (13) and Turkish (15) as compared to the results returned for the search in Greek (4). In contrast, the search results in Greek were distributed more evenly across the different aspects represented, with a larger variety of content / themes.

Furthermore, comparing the results within each aspect, we can see that there is also diversity in the viewpoints represented through each image. For example, the images showing the beaches of Cyprus show a variety of different beaches (e.g., mostly the Nissi beach in the English search results, and the Kyrenia Antique Harbour in the Turkish search results). In fact, different

photographic compositions present the same beach in a different way (e.g., a calm beach, through images without people/with only one person, versus a more lively beach, through images with many people and umbrellas).

Due to the physical division and ongoing political issues of Cyprus, the search results tend to also reflect political perspectives, especially when images of maps and flags are involved. While maps can sometimes be considered as objective representations of a location, maps of Cyprus can highlight the physical divide and embody political commentary. For example, some maps do not show the UN-controlled "Green Line" that runs through the island, dividing it into two, but instead show the districts and geographical attributes on the whole of the island. Others, however, highlight the divide by coloring one half of the island in a deep red color resembling that of blood, and emphasize the political perspective by adding text with politically loaded slogans.

Often, people conduct image searches about a country when they do not have a lot of knowledge about the country and/or are looking to visit the country. Depending on the user's language of choice for the search, they can be presented with dramatically different results for their search, which in turn, can create very different "first impressions" and understandings of Cyprus. When the search is conducted in Greek, e.g., by a Greek-speaking Cypriot or a person from Greece, the results will differ greatly from the results of a search conducted in Turkish, e.g., by a Turkish-speaking Cypriot or a person from Turkey. As the two "sides" of the dispute, these communities are subjected to different extremist points of view, which as we have seen, can be reflected in and reinforced by the search results in their respective languages.

4.2 Google Web search autocomplete suggestions

Autocomplete algorithms, which suggest the next word or phrase for a user to type in the context of a query, represent a form of digital nudging. In spirit, the algorithm aims to assist the user in avoiding spelling mistakes, and in choosing keywords that are likely to lead the user to the desired information. In this regard, suggestions to "autocomplete" someone's query can reveal the most common searches that users conduct related to a topic, including the most frequent questions they ask the search engine. From such questions or statements, we can infer the "reputation" or perception of the search topic, as well as aspects which are investigated most often.

In the context of queries containing the word "Cyprus," the autocomplete suggestions can show us the aspects of Cyprus that presumably, people who are not Cypriots, question the most. Since the English-speaking population is very large and distributed around the world, the autocomplete suggestions in English are assumed to display a global, foreigner's perspective while Greek and Turkish suggestions are assumed to represent Greek and Turkish nationals' perspectives, respectively. The set-up of the examination of the autocomplete algorithm is parallel to that presented in Section 4.1. In other words, we aim to observe the system's behaviour, from the point of view of the same three reference contexts of simulated users, C_G, C_T, and C_E.

Since grammar and syntax differ significantly between the English, Greek, and Turkish languages, we compare the autocomplete suggestions for groups of queries (rather than across individual, translated queries). The queries all refer to either “Cyprus” or “Cypriots”, and often question the reason for an observation that is unfamiliar and/or interesting to the user (e.g., English queries include “why”) such that we can observe the assumptions/perceptions behind the queries prompting the autocomplete suggestions.

English	**Greek**		**Turkish**	
Cyprus				
Cyprus ___	**Κύπρος _____**	**(translation)**	**Kıbrıs ___**	**(translation)**
cyprus mail	Κύπρος πληθυσμός	Cyprus Population	kıbrıs gazetesi	cyprus newspaper
cyprus airways	Κύπρος αξιοθέατα	Cyprus sightseeing	kıbrıs postası	cyprus post [name of a newspaper]
cyprus weather	Κύπρος κατεχόμενα	Cyprus the occupied side	kıbrıs manşet	cyprus headlines
cyprus news	Κύπρος χάρτης	Cyprus map	kıbrıs haber	cyprus news
cyprus times	Κύπρος διακοπές	Cyprus vacation	kıbrıs son dakika	cyprus breaking news
cyprus map	Κύπρος πληθυσμός 2019	Cyprus population 2019	kıbrıs	cyprus
cyprus events	Κύπρος Ελλάδα	Cyprus Greece	kıbrısta ilan	advertisements/listings in cyprus
cyprus league	Κύπρος έκταση	Cyprus Area	kıbrıs hava durumu	cyprus weather
cyprus post tracking	Κύπρος νέα	Cyprus News	kıbrıs iş ilanları	cyprus job listings
cyprus football	Κύπρος 1974	Cyprus 1974	kıbrıs haritası	cyprus map
	Κύπρος ειδήσεις	Cyprus News		
	Κύπρος serial killer	Cyprus serial killer		

Table 4: Autocomplete suggestions for “Cyprus ___” in three languages.

4.2.1 Observations

Similar to the image search results, the autocomplete suggestions for “Cyprus” (in the same three languages) are related to various cultural and political aspects of Cyprus. Again, the different languages can be used as proxies for the nationality (Greek language for those from Greece, Turkish language for those from Turkey) of the users. As observed in Table 4, there is not much commonality in the auto-suggestions across languages; while the number of suggestions provided by the autocomplete algorithms are similar (English and Turkish: 10 each, Greek: 12), the only

themes that appear in all three searches are “news” and “map.” This implies that there is great variety in the types of searches people conduct about Cyprus, in different languages.

In the Turkish results, the news aspect actually takes up the first half (five) of the suggestions, offering specific news outlets as well as different phrasings (all approximately meaning “news”) on which one might search. The English and Greek autocomplete suggestions only have three and two suggestions respectively, by comparison. All three of the English suggestions related to news appear in the first half of the suggestions (and only one is of a specific outlet); the two generic news suggestions for Greek searches appear at places 9 and 11, out of 12. These results show that those searching for Cyprus in Turkish are more likely (as compared to those searching in English or Greek) to be “nudged” towards not just online news, but even to specific news outlets. Of course, this makes it more likely that the perspectives represented in the results for each of these searches are seen - and possibly adopted - by the user.

A search query that simply begins with “Cyprus” might represent what anyone, regardless of cultural or linguistic background, might search on concerning Cyprus. However, people unfamiliar with a location/community may search for questions that they would like to answer, or statements of observations that they find interesting/worth noting. For example, if someone visiting Cyprus observes that people in Cyprus drive on the left, and this is unexpected, they may search for “Why does Cyprus drive on the left.” In order to gauge what aspects of Cyprus are the most unusual to outsiders as well as the assumptions and perceptions foreigners may have, we examine autocomplete suggestions for queries which start with “Why” and follow with either “Cyprus” or “Cypriots”, as well as those that start a statement about “Cypriots”. We use these two categories as guidelines, as the syntax of the three languages do not allow for direct comparison.

In the Annex, the full results from our autocomplete study are presented. As can be observed in those results, the majority of the autocomplete suggestions, regardless of language, discuss the political aspect of Cyprus. These varied in what they referred to as well - e.g. the division/conflict and race (including links to Greece and Turkey), the military, and Cyprus’ links to global politics (e.g., EU, Schengen, NATO, global recognition of the northern part of the island). Making up the majority of our results, these suggestions are likely to interrupt the searcher’s train of thought as they are typing in their query. In other words, instead of following through on an intended search unrelated to politics or the Cyprus Problem, users might be nudged towards searching for political information about Cyprus, which as we have seen, has the potential to be extremist and/or one-sided.

The remaining results presented in the Annex are varied in the aspects of Cyprus that they represented. For example, especially in the English autocomplete suggestions, the queries referred to the geography of Cyprus - e.g., earthquakes, tides, heat, and humidity. Interestingly, many results directly or indirectly commented on the effects of the British colonization of the island, *through other aspects* of Cyprus, e.g., the electrical socket types (infrastructure) or driving on the left (traffic rules). The aforementioned examples highlight aspects that are the most unusual about Cyprus - or at least unusual enough for one to search, to the greatest number of people. The geographical and British aspects were the most common in the English autocomplete suggestions,

while looking closer at the Greek and Turkish autocomplete results revealed questions about name conventions (Greek) and income levels/economy (Turkish). It is clearer, then, to see which aspects of Cyprus differ from the "baseline" perspective of people searching in these languages; i.e., mostly people in Greece and Turkey.

Some rare (but interesting) suggestions were subjective opinions about Cypriots, either as a statement, presumably by those who wanted to find supporting "evidence" for their point of view, or as a question, perhaps by those who were more curious than opinionated (or willing to question a perspective they had come across, which may be "biased"). These suggestions included "Why are Cypriots so loud", "Cypriots are friendly", "Cypriots are lazy", and "Cypriots are a waste of space". All of these, demonstrate the social stereotypes surrounding the Cypriots, which are perpetuated through the autocomplete function of search engines. This observation resonates with previous reports that Google's autocomplete suggestions often reflects prevalent racial stereotypes (e.g., "Why do Black people…") (Baker & Potts, 2013). An example of how one might view their own perspective as "objective" is the suggestion to autocomplete a query to "Why do Cypriots talk differently," where the people searching had the assumption that the Cypriot manner of speaking was simply "different," but without making explicit the point of objective reference.

4.3 YouTube video recommendation engine

When watching a video, there are often recommendations on what to watch next, listed next to the main content. Some people decide not to continue with the video they initially started but instead click one of these videos; others may reach the end of their video and have the first recommendation "autoplay", i.e. start even before they have a chance to react. Therefore, one way or another, video platforms have the opportunity to at least direct people's attention to a relevant video, if not have them watch the whole content. Unfortunately, researchers and journalists have found that these recommendations can "nudge" people towards more extremist views than the first videos they started watching.[8] It has also been found that this nudging is effective; people often watch content that gets increasingly more extremist (especially in regards to politics).[9]

To examine the prevalence and potential effect of the recommendations in the Cyprus context, we search for "Cyprus" in English on the video platform YouTube, and follow the recommendation "trail" down to the fifth recommended video. The methodology, to be precise, is to examine the first result for our query, then five "up next" (UN; first recommendation) videos in a row -- until there are *six* videos in total (five UN). We repeat this process, in a private browser with no cookies and logins, five different times within the same hour.

[8] https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html
[9] https://arxiv.org/pdf/1908.08313.pdf

Video 1	Video 2	Video 3	Video 4	Video 5	Video 6
First search result for "Cyprus"	"Up next" (first recommendation) for Video 1	"Up next" for Video 2	"Up next" for Video 3	"Up next" for Video 4	"Up next" for Video 6

	Trial 1 (T1)	Trial 2 (T2)	Trial 3 (T3)	Trial 4 (T4)	Trial 5 (T5)	Trial 6 (T6)
Video 1 (V1; First search result)	"Cyprus Crisis 1974 - COLD WAR DOCUMENTARY"	[same as T1V1]	[same as T1V1]	[same as T1V1]	[same as T1V1]	[same as T1V1]
Video 2 (V2; "Up Next" for Video 1)	"Six-Day War (1967) - Third Arab-Israeli War DOCUMENTARY"	[same as T1V2]	[same as T1V2]	[same as T1V2]	[same as T1V2]	[same as T4V3]
Video 3 (V3; "Up Next" for Video 2)	"Yom Kippur War 1973: The Egyptian Revenge - (1/4)"	[same as T1V3]	"Russo-Japanese War 1904-1905 - Battle of Tsushima DOCUMENTARY"	"First Arab - Israeli War 1948 - COLD WAR DOCUMENTARY"	[same as T1V3]	"Napoleonic Wars: from Trafalgar to Friedland - Season 1 FULL"
Video 4 (V4; "Up Next" for Video 3)	"Yom Kippur War 1973: The Egyptian Revenge - (2/4)"	[same as T1V4]	"Battle of Königgrätz 1866 - Austro-Prussian War DOCUMENTARY"	[same as T3V3]	[same as T1V4]	"Alexander the Great (All Parts)"
Video 5 (V5; "Up Next" for Video 4)	"Yom Kippur war part 2 - Israel fights for her life and wins"	[same as T1V5]	"Battle of Tuyuti 1866 - War of the Triple Alliance DOCUMENTARY"	[same as T3V4]	[same as T1V5]	"Barbarians - The Saxons"

Video 6 (V6; "Up Next" for Video 5)	"Yom Kippur war part 3 - Israel fights for her life and wins"	[same as T1V6]	"Czechoslovak Legion in Russia and its War to Return Home"	[same as T3V5]	[same as T1V6]	"Barbarians - The Goths"

Table 5: "Up Next" recommendations from YouTube.[10]

4.3.1 Observations

As shown in Table 5, we find that the first search result for the query "Cyprus" (V1) did not change. In fact, Trials 1, 2, and 5 resulted in recommendations for exactly the same videos in the same order. That is, the algorithm performed presumably the same calculations and reached the same result. However, while the procedure remained the same throughout the experiments, the other trials differed. Trials 3 and 4 differed from T1/2/5, but were almost identical to one another, only differing by one video. Trial 6, however, showed us only two videos that had been seen before (V1, as with others, and V2, seen only once before in another trial). Therefore, we can imagine that six people searching for "Cyprus" at the same time will not necessarily be seeing exactly the same videos recommended to them. So, even people searching at the same time from the same place may learn about different aspects of Cyprus, or may see one aspect demonstrated through different perspectives.

As all our trials started from the same video, a documentary about the political conflict in Cyprus about 50 years ago, it is understandable that all recommendations were either about a conflict, historical events/people, and/or countries in the Mediterranean. However, it is interesting to see that in half of our trials, the same "recommendation trail" emerged, discussing a nearby conflict in a similar time period. This trail is made up of multiple parts from two documentaries on the issue, and in the case where these documentaries may favor one "side" of the conflict more than the other, those watching (having initially searched simply for "Cyprus") would be subjected to this perspective. Similarly, two trials follow a similar pattern but include videos on different conflicts around the world, not necessarily in the Mediterranean. Our last trial, notably, does not focus on the conflict aspect much and instead recommends other videos about historical events and people.

5. Conclusions

In this deliverable we presented a diversity framework for analyzing algorithmic system bias. The work is grounded in *perspective-taking*, an approach that advocates that people try to understand one another's positions ("walking in another's shoes"). Given that stakeholders (e.g., user vs. developer) have different reference contexts, this process can promote greater transparency. We have attempted to use this idea in the context of understanding algorithmic system bias, have provided formal definitions (Section 2) and applied those in three different areas for demonstration (text classification, search engines, recommender systems), based on reports from

[10] The videos that appeared multiple times in our experiments are highlighted.

the scientific literature (Section 3). However, taking into consideration the lessons from Standpoint Theory, we have been careful to point out that greater transparency between stakeholders of an algorithmic system, is not sufficient in ensuring that the system is *fair*. As explained, some stakeholders have standpoints that enjoy greater privilege than others in society, which are in line with the prevalent social norms. Therefore, even a transparent system's behaviours need to be evaluated with respect to the specific social context in which the system is deployed, including the power relations between stakeholders within that context, if we are to understand who might be harmed by the system's biases.

Most of the perspective-taking studies discussed in Section 2 involved groups of people coming from different ethic groups, focusing primarily on the American context. Therefore, in order to bring this into perspective with CyCAT's focus, Section 4 examined a number of case studies in information access systems specific to the Cypriot context. The issue of public information access in Cyprus is an interesting case, worth studying mainly due to the political conflict that exists between the two major ethnic communities in the island; Greek-Cypriots and Turkish-Cypriots. The everyday contact we all have with information access systems, allows us to develop a mental model of how these might work. However, the general public is not aware of the personalization that is performed and the different results that each user is consuming. The consequences are even more prominent and are amplified when this is occuring in contexts where political or other conflicts exist.

In our conceptual framework, there are two key roles: i) developers (i.e., anyone who is involved in the development process of a system, and has access to / knowledge of its inner-workings), and ii) observers (i.e., anyone who makes a value judgement - evaluates - the system's behaviours, from "outside," with no access to its inner-workings. Observers might be end-users of the system, but could also be journalists, researchers, auditers, etc.). In that sense, the framework can help us understand potential algorithmic system bias based on perspective-taking. In order to ground our argument for using perspective-taking to understand algorithmic system bias, we presented three case studies of IA systems. IA systems are proprietary and complex in nature, and the developer's context is unknown. We simulated interactions with these systems from the point of view of three user-contexts, using the language of interaction as a proxy for the user's cultural context. Given the outcomes of the case studies we performed, on information access systems, we can demonstrate how these might "nudge" a user in a certain direction in information consumption. Thus, they could be an obstacle to perspective-taking and hence, diversity. These results were consistent in all three cases: image search, text autocomplete, video autoplay list.

More specifically, in the image search results, where the users coming from different parts of the world (according to the language they are using to create search queries) are presented with different images, representing a different impression about the country, e.g., more images depicting a beach came up in English and Turkish compared to the Greek-language search results. Most importantly though, the images reflect in their majority the political divide that exists on the island. When searching in the two non-English languages, the user is prompted to either the Turkish- or the Greek-sided impression of the country's map, reinforcing the user into the

respective extremist view. Revisiting our framework, the system's results are biased towards the one or the other community's reference context and does not allow perspective-taking to develop.

Similarly, in the autocomplete case study users who search using the Turkish language are more likely to be prompted into online political news outlets, compared to the English or Greek speaking users, who perform a search starting with "Cyprus." In all languages, the majority of autocomplete suggestions prompts the user into political information, which emphasizes the political situation on the island. This could be a positive outcome if the results were diverse in all languages and provided information from multiple perspectives. However, the results were different for the two major communities on the island.

When looking at the recommendations on YouTube based on a given video, we can see different trajectories to unroll. Although the trials recommended almost the same series of videos, one can see how one side of the story is presented, forcing a user who is outside of the Cyprus context to develop a one-sided perspective about the political situation in the country.

Given that perspective-taking has been shown to reduce prejudice, bias and stereotyping, and to help people develop empathy when interacting with others within a diverse group, information access systems might also be built in such a way that different perspectives will be promoted. According to our case studies, in the context of Cyprus, users coming from the two major communities, as well as users from outside the island are currently in danger of being led towards an information path that can be biased, promoting extremist political views and one-sided stories.

6. References

Abdollahi, B., & Nasraoui, O. (2018). Transparency in fair machine learning: The case of explainable recommender systems. In Human and Machine Learning (pp. 21-35). Springer, Cham.

Allen, A. (1998). Rethinking Power. Hypatia 13 (1):21 - 40.

Allen, B. (1998). Black Womanhood and Feminist Standpoints. Management Communication Quarterly, Vol. 11, pp. 575-586.

Andreou, A., Venkatadri, G., Goga, O., Gummadi, K. P., Loiseau, P., & Mislove, A. (2018). Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. In NDSS 2018 - Network and Distributed Systems Security Symposium. HAL, San Diego, CA, 1 - 15.

Baker, P., & Potts, A. (2013). 'Why do white people have thin lips?'Google and the perpetuation of stereotypes via auto-complete search forms. Critical Discourse Studies, 10(2), 187-204.

Baeza-Yates, R. (2018). Bias on the Web. Communications of the ACM, 61(6), 54-61.

Barocas, S., Hardt, M. & Narayanan, A. (2018). *Fairness and Machine Learning*. http://www.fairmlbook.org

Calvert, L. M., & Ramsey, V. J. (1996). Speaking as female and white: A non-dominant/dominant group standpoint. Organization, 3(4), 468-485.

Chen, T. W., & Sundar, S. S. (2018, April). This app would like to use your current location to better serve you: Importance of user assent and system transparency in personalized mobile services. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (p. 537). ACM.

Danks, D., & London, A. J. (2017, August). Algorithmic bias in autonomous systems. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (pp. 4691-4697). AAAI Press.

Datta, A., Sen, S., & Zick, Y. (2016, May). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In 2016 IEEE symposium on security and privacy (SP) (pp. 598-617). IEEE.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. Journal of personality and social psychology, 44(1), 113-126.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56-62.

Dignum, V. (2017, August). Responsible autonomy. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (pp. 4698-4704). AAAI Press.

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and mitigating unintended bias in text classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 67-73). ACM.

Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (2017, January). " Be careful; Things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms. In 11th International Conference on Web and Social Media, ICWSM 2017 (pp. 62-71). AAAI Press.

Friedman, B., Kahn, P. H., & Borning, A. (2008). Value sensitive design and information systems. The handbook of information and computer ethics, 69-101.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: decreasing stereotype expression, stereotype accessibility, and in-group favoritism. Journal of personality and social psychology, 78(4), 708-724.

Galinsky, A. D., Todd, A. R., Homan, A. C., Phillips, K. W., Apfelbaum, E. P., Sasaki, S. J., ... & Maddux, W. W. (2015). Maximizing the gains and minimizing the pains of diversity: A policy perspective. Perspectives on Psychological Science, 10(6), 742-748.

Giere, R. (1999). Science without laws. Chicago: University of Chicago Press.

Giunchiglia, F., Maltese, V., & Dutta, B. (2012). Domains and context: first steps towards managing diversity in knowledge. Web Semantics: Science, Services and Agents on the World Wide Web, 12, 53-63.

Gu, L., Yang, P., & Dong, Y. (2017). Diversity optimization for recommendation using improved cover tree. Knowledge-Based Systems, 135, 1-8.

Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., & Wilson, C. (2017, February). Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 1914-1933). ACM.

Hutchinson, B., & Mitchell, M. (2019, January). 50 Years of Test (Un) fairness: Lessons for Machine Learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 49-58). ACM.

Jones, E.E. & Nisbett, R.E. (1971). *The Actor and the Observer: Divergent Perceptions of the Causes of Behavior*. New York: General Learning Press.

Jovchelovitch, S. (2002). Re-thinking the diversity of knowledge: Cognitive polyphasia, belief and representation. Psychologie et société, 5(1), 121-138.

Kay, M., Matuszek, C., & Munson, S. A. (2015, April). Unequal representation and gender stereotypes in image search results for occupations. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 3819-3828). ACM.

Kunaver, M., & Požrl, T. (2017). Diversity in recommender systems–A survey. Knowledge-Based Systems, 123, 154-162.

Kvasny, L. (2006). Let the sisters speak: Understanding information technology from the standpoint of the'other'. ACM SIGMIS Database: the DATABASE for Advances in Information Systems, 37(4), 13-25.

Luke, S. (1974). Power: A Radical View, London: Macmillian.

Mowshowitz, A., & Kawaguchi, A. (2005). Measuring search engine bias. Information processing & management, 41(5), 1193-1205.

Narayanan, A. (2018, February). Translation tutorial: 21 fairness definitions and their politics. In Proc. Conf. Fairness Accountability Transp., New York, USA.

Nelson, B. (2014). The data on diversity. Communications of the ACM, 57(11), 86-95.

Otterbacher, J., Bates, J., & Clough, P. (2017, May). Competent men and warm women: Gender stereotypes and backlash in image search results. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (pp. 6620-6631). ACM.

Otterbacher, J., Checco, A., Demartini, G., & Clough, P. (2018, June). Investigating user perception of gender bias in image search: the role of sexism. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 933-936). ACM.

Rader, E., & Gray, R. (2015, April). Understanding user beliefs about algorithmic curation in the Facebook news feed. In Proceedings of the 33rd annual ACM conference on human factors in computing systems (pp. 173-182). ACM.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59-68). ACM.

Shen, J. H., Fratamico, L., Rahwan, I. & Rush, A. (2018). Darling or Babygirl? Investigating Stylistic Bias in Sentiment Analysis. In Proceedings of 5th Workshop on FAT/ML. Stockholm, Sweden.

Shih, M., Wang, E., Trahan Bucher, A., & Stotzer, R. (2009). perspective-taking: Reducing prejudice towards general outgroups and specific individuals. Group Processes & Intergroup Relations, 12(5), 565-577.

Stoetzler, M., & Yuval-Davis, N. (2002). Standpoint theory, situated knowledge and the situated imagination. Feminist theory, 3(3), 315-333.

Todd, A. R., Bodenhausen, G. V., Richeson, J. A., & Galinsky, A. D. (2011). perspective-taking combats automatic expressions of racial bias. Journal of personality and social psychology, 100(6), 1027-1042.

Tyugu, E. (2006). Understanding knowledge architectures. Knowledge-Based Systems, 19(1), 50-56.

Vescio, T. K., Sechrist, G. B., & Paolucci, M. P. (2003). perspective-taking and prejudice reduction: The mediational role of empathy arousal and situational attributions. European Journal of Social Psychology, 33(4), 455-472.

Appendix

In the table below, the full results from the autocomplete study are presented.

English	Greek		Turkish	
Cyprus				
Cyprus ___	**Κύπρος _____**	**(translation)**	**Kıbrıs ___**	**(translation)**
cyprus mail	Κύπρος πληθυσμός	Cyprus Population	kıbrıs gazetesi	cyprus newspaper
cyprus airways	Κύπρος αξιοθέατα	Cyprus sightseeing	kıbrıs postası	cyprus post [name of a newspaper]
cyprus weather	Κύπρος κατεχόμενα	Cyprus the occupied side	kıbrıs manşet	cyprus headlines
cyprus news	Κύπρος χάρτης	Cyprus map	kıbrıs haber	cyprus news
cyprus times	Κύπρος διακοπές	Cyprus vacation	kıbrıs son dakika	cyprus breaking news [approx]
cyprus map	Κύπρος πληθυσμός 2019	Cyprus population 2019	kıbrıs	cyprus
cyprus events	Κύπρος Ελλάδα	Cyprus Greece	kıbrıs'ta ilan	advertisements/listings in cyprus
cyprus league	Κύπρος έκταση	Cyprus Area	kıbrıs hava durumu	cyprus weather
cyprus post tracking	Κύπρος νέα	Cyprus News	kıbrıs iş ilanları	cyprus job listings
cyprus football	Κύπρος 1974	Cyprus 1974	kıbrıs haritası	cyprus map
	Κύπρος ειδήσεις	Cyprus News		
	Κύπρος serial killer	Cyprus serial killer		
"Why" + Cyprus				
why is cyprus ___	**Γιατί η Κύπρος _____**	**(translation)**	**neden Kıbrıs ___**	**(translation)**
why is cyprus so hot	Γιατί η Κύπρος είναι Ελληνική	Why is Cyprus Greek	neden kıbrıs tatlısı	why cyprus dessert
why is cyprus so expensive	Γιατί η Κύπρος ονομάστηκε έτσι	Why was Cyprus named like this	neden kıbrıs'a giden yunanistana giremiyor	why can people who go to cyprus not get into greece
why is cyprus so british	Γιατί χάθηκε η Κύπρος	Why was Cyprus lost	neden kıbrıs	why cyprus

why is cyprus so humid	Ορέστης Κύπρος γιατί σκότωνε	Orestes Cyprus why did he kill	kıbrıs neden türkiye'ye bağlanmıyor	why doesn't cyprus connect to turkey
why is cyprus divided	Γιατί η Κύπρος δεν είναι Ελληνική	Why is Cyprus not Greek	kıbrıs neden önemli	why is cyprus important
why is cyprus so windy	Γιατί η Κύπρος δεν έχει ναυτικό	Why doesn't Cyprus have a navy	kıbrıs'a neden yavru vatan deniyor	why is cyprus called "child land"
why is cyprus in the eu			kıbrıs neden türkiye'ye katılmıyor	why doesn't cyprus join turkey
why is cyprus not in nato			kıbrıs neden ingilizlere verildi	why was cyprus given to the british
why is cyprus still divided			kıbrıs neden tanınmıyor	why is cyprus not recognized
why is cyprus not in schengen			kıbrıs neden ikiye ayrıldı	why did cyprus divide into two
why does cyprus ___	**Γιατί η Κύπρος είναι ____**	**(translation)**	**Kıbrıs neden ___**	**(translation)**
why does cyprus speak greek	Γιατί η Κύπρος είναι Ελληνική	Why is Cyprus Greek	kıbrıs neden türkiye'ye bağlanmıyor	why won't cyprus connect to turkey
why does cyprus have english plugs	Γιατί η Κύπρος δεν είναι Ελληνική	Why is Cyprus not Greek	kıbrıs neden tanınmıyor	why is cyprus not recognized
why does cyprus drive on the left	Γιατί η Κύπρος δεν είναι στο νατο	Why is Cyprus not in NATO	kıbrıs neden önemli	why is cyprus important
why does cyprus have so many cats	Γιατί η Κύπρος είναι θύελλα	Why is Cyprus a storm	kıbrıs'a neden yavru vatan deniyor	why is cyprus called "child land"
why does cyprus not have tides	Γιατί η Κύπρος είναι σεισμογενής χώρα	Why is Cyprus an earthquake-prone country	kıbrıs neden türkiye'ye katılmıyor	why doesn't cyprus join turkey
why does cyprus have so many guns	Γιατί η Κύπρος είναι δική σου	Why is Cyprus yours?	kıbrıs neden ingilizlere verildi	why was cyprus given to the british
why does cyprus have a buffer zone			kıbrıs neden ikiye ayrıldı	why did cyprus divide into two

why does cyprus have earthquakes			kıbrıs neden bölündü	why was cyprus divided
why does cyprus use the euro			kıbrıs neden önemli kısaca	why is cyprus important [explained] shortly
why do cyprus drive on the left			kıbrıs neden pahalı	why is cyprus expensive
"Why" + Cypriots				
why are Cypriots ____	**Γιατί οι Κύπριοι ____**	**(translation)**	**neden Kıbrıslılar ___**	**(translation)**
why are cypriots so dark	Γιατί οι Κύπριοι μισούν τους Έλληνες	Why do Cypriots hate Greeks	kıbrıslılar neden türkleri sevmez	why don't cypriots like turks
why are cypriots so loud	Γιατί οι Κύπριοι μιλάνε Ελληνικά	Why do Cypriots speak Greek	kıbrıslılar neden türkiyeyi sevmez	why don't cypriots like turkey
why are cypriots called charlies	Γιατί οι Κύπριοι είναι Έλληνες	Why are Cypriots Greeks	kıbrıslılar neden zengin	why are cypriots rich
why are cypriots greek	Γιατί οι Κύπριοι έχουν το ίδιο όνομα και επίθετο	Why do Cypriots have the same name and surname	kıbrıslılar neden farklı konuşur	why do cypriots talk differently
why are english cypriots called charlie			kibrislilar neden türkleri sevmez	why don't cypriots like turks
why turkish cypriots are				
why do Cypriots ___	**Γιατί οι Κύπριοι είναι _____**	**(translation)**	**Kıbrıslılar neden ___**	**(translation)**
why do cypriots speak greek	Γιατί οι Κύπριοι είναι Έλληνες	Why are Cypriots Greek?	kıbrıslılar neden türkleri sevmez	why don't cypriots like turks
why do cypriots drive on the left			kıbrıslılar neden türkiyeyi sevmez	why don't cypriots like turkey
why do cypriots smash plates			kıbrıslılar neden zengin	why are cypriots rich
why do cypriots call themselves greek			kıbrıslılar neden farklı konuşur	why do cypriots talk differently
why do cypriots say re			kibrislilar neden türkleri	why don't cypriots like turks

			sevmez	
what do cypriots look like			kıbrıslılar türkiyelilere neden kabasakal der	why do cypriots call turks "kabasakal" [a slang word]
what do cypriots eat			kıbrıslılar türkleri neden sevmiyor	why don't cypriots like turks
what do cypriots think of the british			kıbrıslılar türk askerini neden sevmez	why don't cypriots like the turkish soldiers/army
what do cypriots speak				
what do cypriots think of othello				
Cypriots				
are Cypriots ___	**Κύπριοι ___**	**(translation)**	**Kıbrıslılar ___**	**(translation)**
are cypriots white	Κύπριοι ευρωβουλευτές 2019	Cypriot MEPS 2019	kıbrıslılar birliği	cypriots union
are cypriots greek	Κύπριοι ευρωβουλευτές	Cypriot MEPs	kıbrıslılar türk mü	are cypriots turkish
are cypriots middle eastern	Κύπριοι ποδοσφαιριστές	Cypriot football players	kıbrıslılar nasıl konuşur	how do cypriots speak
are cypriots friendly	Κύπριοι συνθέτες	Cypriot composers	kıbrıslılar neden türkleri sevmez	why do cypriots not like turks
are cypriots caucasian	Κύπριοι influencers	Cypriot influencers	kıbrıslılar ingiliz vatandaşı	cypriots british citizens
are cypriots muslim	Κύπριοι γλύπτες	Cypriot sculptors	kıbrıslılar vietnamda	cyoriots in vietnam
are cypriots arab	Κύπριοι ολυμπιονίκες	Cypriot Olympians	kıbrıslılar nasıl insanlardır	what kind of people are cypriots
are cypriots asian	Κύπριοι ήρωες	Cypriot heroes	kıbrıslılar yalısı	cypriot waterside mansion [proper noun]
are cypriots greek or turkish	Κύπριοι δημοσιογράφοι	Cypriot journalists	kıbrıslılar ekşi sözlük	cypriots sourdictionary [name of a turkish forum like reddit]

	Κύπριοι ποιητές	Cypriot poets	kıbrıslılar türkleri neden sevmiyor	why do cypriots not like turks
Cypriots are ___				
cypriots are greek				
cypriots are not greek				
cypriots are lazy				
cypriots are awesome				
cypriots are a waste of space				
are cypriots white				
are cypriots arab				
are cypriots caucasian				
are cypriots muslim				
are cypriots asian				