Diversity by Design? Balancing the Inclusion and Protection of Users in an Online Social Platform

Paula Helm*
paula.helm@tum.de
Technical University of Munich
Munich, Germany

Loizos Michael*
loizos@ouc.ac.cy
Open University of Cyprus &
CYENS Center of Excellence
Nicosia, Cyprus

Laura Schelenz*
laura.schelenz@uni-tuebingen.de
University of Tuebingen
Tuebingen, Germany

ABSTRACT

The unreflected promotion of diversity as a value in social interactions — including in technology-mediated ones — risks emphasizing the benefits of inclusion without recognizing the potential harm of failing to protect vulnerable individuals or account for the empowerment of marginalized groups. Adopting the position that technology is not value-neutral, we seek to answer the question of how technology-mediated social platforms can accommodate *diversity by design* by balancing the often tension-ridden principles of protection and inclusion. In this paper, we present our research program, developed strategy, as well as first analyses and results. Building on approaches from scenario analysis and Value Sensitive Design, we identify key arguments for a "diversity by design"-agenda. Furthermore, we discuss how these arguments can be operationalized and implemented in a diversity-aware chatbot and provide a critical reflection on the limits and drawbacks of the proposed approach.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Scenario-based design**; *Collaborative and social computing*; • **Social and professional topics** → *Hate speech.*

KEYWORDS

diversity, ethics, content moderation, social media, platform, matching algorithms, design, norms, intervention, inclusion, protection

ACM Reference Format:
Paula Helm, Loizos Michael, and Laura Schelenz. 2022. Diversity by Design? Balancing the Inclusion and Protection of Users in an Online Social Platform. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3514094.3534149

1 INTRODUCTION

Diversity refers to the difference of "things" such as opinions, cultural backgrounds, socio-economic statuses, and routine practices [2, 46]. Yet, it is also associated with different values such as inclusion, tolerance, and equality [59, 63]. In the computing technology industry, diversity increasingly gains attention, often as a desideratum for the design of digital technologies [8, 49]. Technology, and new media technology in particular, indeed enables and is enabled by diversity as it connects people with different skills, experiences, and opinions across geographic and cultural boundaries [7].

At the same time, research points to the risks that could result from the machine-mediated interaction of people with different experiences and opinions, including hate speech and gendered or racial violence [3, 31], the spread of "fake" or "junk news" [41, 45, 60] as well as algorithmic discrimination [14, 43]. Thus, if it comes without appropriate protection measures, diversity as a desideratum for online communities can hinder rather than foster equitable engagement and exchange. Ethics-driven designers and platform operators therefore face a "diversity dilemma". They must balance the inclusion of people with diverse experiences and opinions with the protection of marginalized and / or vulnerable users [21]. Seemingly paradoxically, leveraging diversity may require limiting diversity.

In this paper, we start from the premise that one of the key challenges of diversity by design is to balance the goal of maximum inclusion with the need to protect vulnerable groups. While freedom of expression is the desirable default state and requires the inclusion of as much diverse content, actors, and forms of communication as possible, there is also a need to create safe spaces to enable protected communication, especially around sensitive topics and actors [27]. However, there is also a risk of overly restricting diversity, and if restricting access to communication for the sake of protection is not ethically legitimate, it may amount to outright discrimination. We explore strategies for addressing the challenge of balancing inclusion and protection in an online social platform. We argue that the most effective, appropriate, secure, and ethical solution to these challenges is a semi-automated, distributed responsibility system in which designers and developers share responsibility for balancing protection and inclusion with a system's users.

Our work is related to recent discussions about appropriate content moderation strategies in online social networks [22, 23, 48]. However, we deliberately chose to use the term "curation" to emphasize the explicitly ethics-driven impetus of our work and because we view diversity as a category that is even more multifaceted and pervasive than content. By curation, we do not mean the management or exhibition of diversity. Rather, we take the term literally, recalling that curation in its Latin origin descends from the word care, and by care we mean, according to Fischer and Tronto's famous definition, "everything we do to maintain, preserve, and repair 'our world' so that we can live in it as best we can" [16].

*All authors contributed equally to this research.

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534149

Approaching diversity as both a descriptive and a normative category, in this paper, we adopt an interdisciplinary approach that combines Ethics and Computer Science with perspectives from Science and Technology Studies. Our research is based on the Values in Design (VID) paradigm, which highlights the fact that technology is always value-laden, and in most cases reflects the value hierarchies of its designers, developers, clients, and commissioners [17, 19, 29]. For our empirical analysis, we combine the VID methodology with approaches from the field of scenario analysis [12, 50], which allows us to evaluate real-world scenarios that have already emerged in the testing phase of our use-case. The use-case that we consider is a Telegram-based chatbot that allows users to ask and answer questions to a community based on the WeNet online social platform. The chatbot and the platform were developed as research tools in the context of the WeNet research and innovation project [64].

The question guiding our analysis and discussion is the following: *How can we build interfaces and algorithms that support meaningful user interactions in online social platforms in a way that considers the values and risks of diversity?* More specifically geared towards our use-case and the encoding of norms into an online social platform, we ask: *What norms can we formalize and encode into our online social platform to ensure that it satisfies our ethical responsibilities for both the inclusion and protection of users?* We follow a three-step strategy: *(1)* drawing on ethical reasoning, we identify the values and contexts that are relevant to the curation of diversity in a platform-based online community, *(2)* working with real-world scenarios, we determine norms that can be formalized to guide the machine-mediated balancing of inclusion and protection, *(3)* in the implementation, we consider how to develop actionable policies for the machine-mediated balancing of inclusion and protection.

The paper makes the following contributions: In Section 2, we clarify our understanding of diversity and its implications in the context of an online social platform. In Section 3, we propose a three-step strategy to determine how we can balance protection and inclusion. In the first, theory-driven, step (cf. Section 4), we develop a framework for identifying relevant values and the social spheres that contextualize the values. This serves as the basis for the second, scenario analysis, step (cf. Section 5), from which we derive four norms that guide ethical intervention in terms of limiting or expanding access to communication based on the users' preferences. Finally, in the third, operationalization, step (cf. Section 6), we present and discuss different options for implementing ethically-inspired interventions, including automation, machine learning, crowd-sourcing, natural language processing, and design choices.

2 ON THE MEANING OF "DIVERSITY"

We start by explaining how we understand diversity in this paper, and its relation to the need to balance protection and inclusion.

2.1 A Normative and Conceptual Definition

In public discourses, "diversity" is commonly framed as something "good", invoking associations with pluralism, tolerance, and inclusion [61]. Diversity has also become a popular marketing strategy and has found its way into the rhetoric of large technology companies [8]. In the process, diversity is often conceptualized and treated as a kind of resource that can be "exploited". However, diversity is more complicated. Below, we differentiate between normative and conceptual diversity to better distinguish values associated with diversity and the actual notions of difference that usually underlie the designers' and the developers' understandings of diversity.

From an ethical perspective, diversity can be seen as a value that is either intrinsic or instrumental. In the former case, diversity is good by and for itself. In the latter case, diversity helps achieve other values such as inclusion, tolerance, equality, and democracy, and vice versa [66]. Beyond the instrumental value of diversity for a democratic and just society, diversity can serve to broaden individuals' experiences and knowledge. For example, exposure to different lifestyles can enable individuals to better empathize with the realities of other people's lives. Exposure to cultural diversity can ignite curiosity as a driver of progress and knowledge. The UNESCO "Convention on the Protection and Promotion of the Diversity of Cultural Expressions" also supports the idea of cultural diversity as a means of promoting tolerance and mutual understanding [59]. Finally, diversity can help advance organizational goals, such as solving complex tasks, by bringing diverse team members together in productive ways [2]. Diversity is, hence, a constitutive feature of values and moral practices, making it ethically important.

Diversity can also be defined as a descriptive tool. Such an understanding of diversity is widely used in various fields of study. As a descriptive category, diversity helps us define the differences between users and ultimately classify elements of their personalities (e.g., skills and practices) that can complement the characteristics of other users to create a moment of mutual support [26]. The Telegram-based chatbot that serves as our use-case was developed as part of a project in which a descriptive understanding of diversity (a diversity model) was put to use. This model is based on different social practices and demographic characteristics of users [64]. This kind of operationalization of differences between users, which is built not only on demographics but also on practices forms the basis for matching users in the community according to their interests, psychological make-up, character traits, skills, and needs.

In this paper, we build on this descriptive model of diversity, but take it a step further by seeking an ethics-oriented curation of diversity in terms of balancing protection and inclusion. For this purpose, we combine the normative and the descriptive perspective of diversity because first, we need a goal that guides the ethical interventions in the online social platform, and second, we need to operationalize and represent diversity in order to protect it.

2.2 The Protection / Inclusion Tension

As an instrumental value, diversity can help achieve other important values. However, in order to unleash diversity's benefits, diversity itself depends on the configuration of other instrumental values. In a simplified representation, we can argue that protection and inclusion are constituent features of diversity, and vice versa (cf. Figure 1). However, protection and inclusion can be in tension with each other, as in the case of online social platforms, requiring a balancing act to accommodate these two aspects of diversity.

The inclusion of as many different perspectives as possible in an online social platform can help users explore different views. Maximum diversity or exposure to "other" opinions can be instrumental in breaking the much-cited "filter bubble" [44]. While the term itself

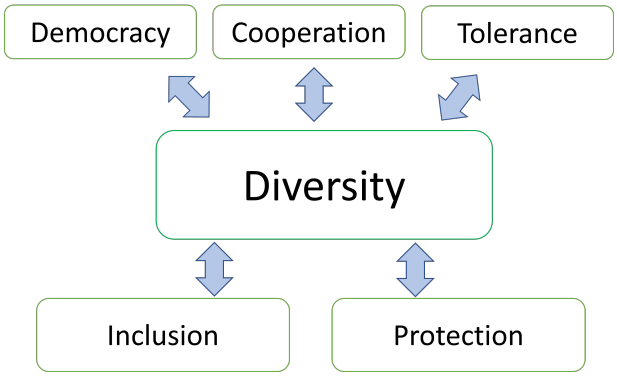

Figure 1: The relation of diversity to other values, when diversity is approached as an instrumental value.

is contested and different authors may refer to different things, the general idea is that users are trapped in a chamber echoing their own beliefs and interests [6]. The problematic result is a lack of exchange with people who exhibit diverging interests and opinions.

In the realm of commercially-driven algorithmic mediation, echo chambers quickly emerge from the logic of maximum user engagement, because a computer model is optimized to make personalized recommendations based on the target users' previous interests [1]. Algorithmic recommendations thus risk encouraging users to consume similar interests or relationships. Algorithmic echo chambers should be mitigated to promote diversity because users' engagement with diverse interests and relationships is associated with their increased cultural sensitivity, and is essential for democratic ideals [25, 59]. Furthermore, exposure to different perspectives and skills is said to improve the performance of users in work contexts, an argument that is prevalent in (the technology) industry [2, 8].

However, these benefits of diversity must be protected. Hate speech and other derogatory forms of communication and behavior can cause harm in general. But they are most destructive when power imbalances are at play. This takes place where, for example, marginalized or otherwise vulnerable groups of people are inadequately protected. When diversity of opinion is unregulated, racist, ableist, and sexist content can harm the rights and freedoms of minorities [11, 31, 34]. Supremacist content can not only insult and question the humanity of minorities, but also slowly erode the foundation of liberal values [3, 11]. Moreover, blatant hateful attacks against minorities on social media have a silencing effect, as those attacked fear for their lives or safety if they speak out [9, 34].

Given these risks, providing safe spaces or other forms of closed-communication areas in an online social platform is crucial, as is regulating language or moderating content [22]. Although this may sound counter-intuitive, limiting diversity in one respect can sometimes promote diversity in another [62]. Two contexts can be highlighted here: The first is a political one, in which identity groups join with their peers to discuss oppression and develop social justice campaigns. The second is a health-related one, in which sharing experiences and coping with the difficulties of managing illness requires the understanding and empathy of peers [27]. Seemingly paradoxically, therefore, safe spaces with restricted access to certain groups of people may turn out to be a means of promoting inclusion.

3 METHODOLOGICAL APPROACH: VALUE-CENTERED SCENARIO ANALYSIS

In our attempt to tackle the challenge of curating diversity in an online social platform through balancing protection and inclusion, we follow the Values in Design research program [9, 29]. This research and design program departs from a widely-held understanding of technology as being objective and value-neutral. Value Sensitive Design (VSD) is a particularly pragmatic approach in this regard because it does not merely reflect on the values that are consciously or unconsciously inscribed in technology. Instead, VSD goes one step further and proactively seeks to identify key norms to guide the inscription of values into technologies. Moreover, VSD evaluates the potential range of tools that may be employed to realize this process, and it reflects on their future implications [19].

In our research, we combine VSD with approaches from the field of scenario analysis [12, 50]. Scenario analysis is a method originating from the field of Futurology that is applied to envision multiple possible futures and prepare and develop ways to deal with them appropriately. Value scenarios are narratives that tell stories about use and are intended to show the entanglement of human and technical aspects in specific contexts. They emphasize impacts on direct and indirect stakeholders, associated key values, typical uses, indirect impacts, and similar systemic effects [42].

Bringing together the aforementioned theoretical lenses and methods, in the following we propose a strategy for determining how to best (in the sense of most effectively *and* ethically) curate diversity through the balancing of protection and inclusion. Our proposed and adopted strategy consists of the following three steps:

(1) Assess the relevant values, and the factors that contextualize them, that affect the making of an ethically-informed decision on how to balance protection and inclusion.

(2) Empirically determine how to balance the relevant factors through a value-centered scenario analysis of exemplary cases and identify the key ethical norms that apply.

(3) Technically implement the identified ethical norms into an actionable policy by operationalizing the relevant factors in a formal representation that can support the taking of actions based on the prescriptions of the ethical norms.

The three steps of the strategy are not meant to be understood as being invoked sequentially and once-off. Rather, they tend to be visited iteratively and possibly out of order until convergence.

Before we expand on our work within each step, we introduce the use-case on which the strategy has been applied. Although not all steps of the strategy are necessarily dependent on the details of the given case, we nonetheless consider it instructive to present the case upfront, so that the reader can mentally ground our analysis.

3.1 Use-Case: A Telegram-Based Chatbot

Our considered use-case involves the online social platform WeNet, where users, designers, and researchers come together as a community to develop services, build apps, and conduct research with data collected through those apps. In this paper, we focus our use-case

on questions revolving around the design and implementation of the chatbot Ask4Help, developed as a research tool by a consortium of European and non-European partners from more than 10 countries [64], and tested on pilot sites in Northern and Southern Europe, Latin America, as well as East and Central Asia.

The chatbot is designed as an instant messaging application within the Telegram instant messenger, running on the diversity-aware platform WeNet. Telegram was chosen for the following reasons: *(a)* it supports the implementation of chatbots, *(b)* participants are provided with software updates making the chatbot compatible with different hardware and operating systems, *(c)* users have the option to keep their personal data largely undisclosed, and *(d)* the implementation effort was minimal, requiring work only on the back end, without worrying about the user-facing side.

The chatbot allows users to send questions (or "requests") to the chatbot's community of registered users. When a user submits a question to the community via the "/question" command, this is matched to candidate respondents by a WeNet platform algorithm, who are then given the opportunity to respond to the question. The matching process is based on continuously updated and enriched user profiles to identify the most appropriate respondents.

The user profiles are populated with data collected through self-report, chat activity, mobile sensor and geolocation tracking. Here, our conceptual understanding of diversity comes into play (cf. Section 2.1), which implies not only demographic data, but also social practices, that is, skills, interests, habits, and psychological make-up. Such a classification goes beyond traditional approaches to user profiling. Although the operationalization is not without flaws and ethical concerns, its major advantage is to capture not only "hard" but also "soft" differences among users. In a nutshell, diversity in the chatbot is understood as the different demographic attributes *and* social practices enacted by the members of a community.

When selecting the set of candidate respondents, an algorithm could, in principle, employ one of the following default baseline matching criteria (in addition to any other considerations):

Only Similar: the system sends requests only to candidate respondents who are similar to the user, either in terms of demographic attributes or social practices or both.

Only Different: the system sends requests only to candidate respondents who are different from the user, either in terms of demographic attributes or social practices or both.

No Restriction: the system sends requests to candidate respondents that are either similar to or different from the user, and is, thus, most inclusive of community members.

In the chatbot, these hypothetical defaults do not materialize because it is left to the users to indicate whether they wish to ask users who are "similar" or "different" (the criteria for similarity or difference vary in terms of the questions being asked and can relate to demographics, world-views, skills, etc.). The availability of this option to the users was originally intended for research purposes to gain a better understanding of users' needs, but also allows users to affect the set of candidate respondents in relation to the question.

From an ethical perspective, it is important to empower users with increased control over algorithmically-mediated communication. This said, our analysis in Section 5 will consider whether it is possible and ethically-legitimate to intervene in the users' preferences for diversity, e.g., when users (intentionally or unintentionally) exclude marginalized groups with their particular choice.

When matching users' requests, the system relies on two streams of information: *(i)* individual preferences for diversity (Does the user require someone with similar experiences / characteristics to create a safe space, or can they benefit from expertise / skills that complements their own?), and *(ii)* demographic diversity (Who is in the position to provide a useful response to the user's question?).

The work presented in this paper — aiming to culminate in an ethical intervention policy to manage protection and inclusion of platform users — was initiated after a first round of pilot studies with the chatbot showed that users regularly asked highly sensitive questions. These related to issues concerning mental health (e.g., exam anxiety), suicidal fantasy and isolation, as well as politically controversial topics related to the COVID-19 pandemic, and questions about spirituality and marriage (which may not be considered sensitive by members of liberal cultural communities, but fall under this category when addressed by members of more conservative ones). To mitigate the potential harm faced by users who raise sensitive topics through the WeNet platform, the ethics and design team decided to explore additional algorithmic measures and design tools to provide necessary protections. In the process, the difficulty became apparent that while protection is necessary, it must be carefully balanced with the principle of inclusion.[1]

4 STEP 1: THEORY-DRIVEN ASSESSMENT OF RELEVANT VALUES AND CONTEXTS

The first step of our adopted strategy draws on work from the field of Political Philosophy to provide a robust theoretical foundation on which we can, in the second step, assess the relevant norms at stake in a given scenario. As laid out in Section 2, diversity relates to other values in dynamic ways. Given the multiplicity of values involved in these complex constellations, value tensions might arise. Our assessment brings together both instrumental and intrinsic or fundamental values (cf. [66]) in order to build the foundation on which ethical reasoning about the curation of diversity is possible.

Agreeing on a positive notion of diversity is a characteristic of modern liberal societies. Relevant instrumental values with respect to diversity are: tolerance, freedom of choice, efficiency, inclusion, and protection. According to Forst [18], instrumental values need to be evaluated as normatively dependent to the extent to which they contribute to the promotion of fundamental values. Fundamental values according to the European Union, whose perspective we adopt here, are: dignity, freedom, democracy, equality, rule of law, and human rights [15]. Fundamental values should not be violated and thus, if there are value tensions in an interaction between users in an online social platform, the design should take into consideration the severe consequences of constraining fundamental values. Like instrumental values, fundamental values do not exist in isolation, but are tied to specific socio-historical and cultural contexts. In order to take these contexts into account, in addition to

[1] The spread of misinformation is also an issue that accompanies online diversity, and which needs to be part of a comprehensive curation process. This, however, exceeds the scope of this paper, in which we focus on strategies to balance protection and inclusion with regard to hate speech, silencing and discrimination, more specifically.

Figure 2: The use-case technology as a chatbot supporting machine-mediated human-to-human interactions.

the differentiation of instrumental and fundamental values, we also introduce the idea of different social spheres as context factors.

Inspired by Walzer [63] and influenced by Taylor [54], we consider the social contextuality of value judgments depending on the social spheres at stake in each scenario. This is important when dealing with scenarios where we wish to balance different fundamental or different instrumental values that are in tension with each other. For this purpose, we make an ideal-type distinction between three spheres. We deliberately call this distinction "ideal-typical" because we acknowledge that these spheres are not given, but are socially constructed, contingent, and indeed fluid; they represent "social imaginaries" of modern social forms [54]. Social imaginaries are, to use Charles Taylor's term, the embodied, imagined, and narrated ways in which large groups of people in a society imagine their social existence, the moral order on which it is based, how they fit in with others, the expectations that are usually met, and the norms associated with them. Social imaginaries are thus carriers of shared understanding, but they are not just in our heads. Rather, they shape social practices through which a widely shared sense of legitimacy is enabled [54, pp. 23–29]. As such, they matter.

According to Taylor [55], the modern social imaginary, which is part of the basic structure of Western democracies, comprises a system of interlocking spheres, which include: *(1)* the private sphere, in which intimate and sensitive issues have their place, *(2)* reflexivity, commons and the social contract, public opinion and the public sphere of democratic deliberation, and *(3)* the market economy as an independent force and the self-government of citizens within a society as a normative ideal [55]. Correspondingly, we distinguish between, and adopt, the following three spheres: *(1)* sensitive sphere, *(2)* public sphere, and *(3)* self-governance sphere.

The idea of structuring societies through differentiation between spheres is certainly not without flaws, yet, it is helpful for identifying public policies or clarifying our expectations towards different social contexts. For instance, in a strongly condensed way, it can be stated that in the context of data collection under the EU General Data Protection Regulation, the rules for data processing are stricter when it comes to the sensitive medical sphere, where the protection of privacy is paramount. This is in contrast to the public political sphere, where transparency is generally more important [56].

At this point, it should be noted that when we refer to Taylor's concept of social imaginaries, we are talking about Western imaginaries. This means that this part of our strategy is specific to the Western cultural context, or more precisely to the European one, since we also adopt the core values of the European Union. Accordingly, our framework is not readily transferable to other cultural contexts without friction [58]. If necessary, this step of our strategy must be adapted to the cultural context in which it is applied.

Moreover, the idea of different spheres brings its own problems and has been criticized many times, e.g., by feminist and other human rights movements [10]. For example, it has been argued that the simplistic distinction between a private and a public sphere can be misused as a shield to inspect, oppress, and silence women, to cover up abuses, and to obscure historically established structures of domination [33]. We treat these spheres not as natural categories, but as historically rooted and performative social imaginaries that continue to shape social practices in Western societies today. This allows us to acknowledge power relations, while having a pragmatic guide to make actionable ethical and socio-culturally grounded decisions about diversity curation in an online social platform.

5 STEP 2: EMPIRICAL DETERMINATION OF NORMS FOR DIVERSITY CURATION

The second step entails a scenario analysis through which we generate the relevant norms required for the third step. Here, we consider different scenarios where users have made a request to "someone similar" or "someone different". In each of the scenarios, we analyze the values and the social spheres involved in the interaction, and relate them to needs for inclusion and / or protection (cf. Section 2.2). Apart from sensitive issues, sometimes, a question may require a group of "similar" people to answer it due to demographic and efficiency-based reasons. For instance, when a pregnant user asks about challenges during pregnancy, they likely benefit from responses from other pregnant users or those who have been pregnant. In other instances, users may benefit from respondents who

Table 1: Overview of scenarios and the values and spheres that apply. Dominant values that trump others are shown in bold.

scenario #	Scenario 1	Scenario 2	Scenario 3	Scenario 4
user profile	Khulan. Suffers from test anxiety. On a scholarship at a university abroad.	Sanaya. Exchange student from Bangladesh at University of Trento, Italy. Into photography.	Sanaya. Exchange student from Bangladesh at University of Trento, Italy. Into photography.	Daniel. Studies German language and literature in Germany. Writer, has written a novella.
question	"Are you struggling with mental health issues that can make it difficult for you to study? And if so, how do you deal with them?"	"What kind of camera is a good one to buy for landscape photography?"	"Are there other semi-professional photographers interested in starting a photography group?"	"I am looking for a native German speaker to edit a 90-page literary text for €500. Interested?"
similar/different	someone "similar to me"	someone "similar to me"	someone "similar to me"	someone "similar to me"
values involved	well-being, health, **dignity**	**efficiency**, freedom of choice	inclusion social justice, freedom of choice, **autonomy**	freedom of choice, efficiency, autonomy, inclusion, **equality**
sphere (context)	sensitive sphere	self-governance sphere	self-governance sphere	public sphere
applicable norm	safe-space norm	efficiency norm	freedom-of-choice norm	non-discrimination norm
design action	limit diversity	limit diversity	limit diversity	keep or expand diversity

are decidedly different with regard to the question at hand, e.g., because their competences complement those of the user.

The chatbot's feature to chose "someone similar" or "someone different" is meant to increase user agency. But it may also trigger discriminatory effects resulting from an overuse of the option "someone similar". For instance, if a request can be relevant to a broader group of people but is constantly shared among the same group, some users are excluded from relevant opportunities. This is especially pertinent with questions that are linked to the public sphere. In such cases, we may consider to intervene with design choices (cf. Section 6.3) that remind users of the value of inclusion. Reversely, if a request is sensitive and likely provokes offensive reactions, it may be advisable to limit the group of people either to peers who are familiar with the issue or to professionals. In such a case, we again may consider serving the user with a soft intervention that triggers reflection on the scope of exposure.

In order to make design choices, we must first determine the norms that direct curation. To this end, we ask: *In a given scenario, is it ethically legitimate that the user restricts or limits the group of recipients by choosing to send a request to "someone similar"?* The following scenarios help us arrive at an ethically and empirically informed decision about the appropriate action in terms of diversity.

5.1 Exemplary Real-World Scenarios

In the scenario analysis we conduct based on our proposed strategy, we focus on four settings in which arguments for inclusion and protection are in tension. In three settings, different instrumental and fundamental values are weighted against the inclusion and protection arguments. Here, it is legitimate to limit diversity despite limited inclusion, that is, to choose the option to send a request to "someone similar". Additionally, we present one setting where the fundamental value of equality is at stake, trumping all instrumental values, thus making it illegitimate to limit diversity due to discriminatory effects. This requires either choosing the option to send a request to "someone different", or expressing no preference (which effectively sends the request to both those who are similar *and* those who are different). An overview of the settings and the values and spheres involved in each setting is given in Table 1.

Scenario 1: Asking for psychological advice.

Setting: Khulan is studying at a university far from her country. The student suffers from test anxiety. Starting with mild symptoms, the anxiety grew progressively worse. Khulan began skipping exams, lagging behind in the curriculum. Seeking peer support, she turns to the WeNet platform. She asks through the chatbot, "Are you struggling with mental health issues that can make it difficult for you to study? And if so, how do you deal with them?". Khulan chooses to ask similar people (e.g., people who also suffer from test anxiety). She does so to make sure that she is not exposed to competitive peers who find studying easy and whose potentially insensitive comments would increase her feelings of failure.

Ethical reasoning and design action: Limit diversity, for protection and to promote dignity. In cases related to (mental) health issues, it is not only legitimate but even recommended for individuals to choose a restricted target group for sharing their concern. In medical / psychological contexts, protection is paramount to prevent stigma, discrimination or otherwise damaging reactions. A nuanced limitation of diversity can empower users to create a safe and protected space for discussing their concern with peers.

Scenario 2: Buying a camera.

Setting: Sanaya, a student from Bangladesh, begins her MA of Arts at an Italian university. She is fascinated by photography as well as plants and animals. Wanting to capture the landscape, she finds that her camera is unsuitable for the purpose. She asks peers through the chatbot: "What kind of camera is a good one to buy for landscape photography?". She decides to ask people who are similar to her (e.g., people who are also into photography), as she seeks advice from people with relevant experience.

Ethical reasoning and design action: Limit diversity, for efficiency reasons. Limiting diversity here is intuitive in order to find the most useful answer. It is reasonable to assume that someone who already owns a camera, or is interested in photography, can answer the request well. Here, limiting diversity is done not due to the sensitivity of the question but rather due to efficiency. Efficiency is, ethically speaking, not as strong a case as protection. But the case at hand merely relates to leisure activities, the self-governance

Figure 3: Illustration of Scenario 1 from Section 5.1, with a student asking a sensitive question via the chatbot.

sphere. It does not include a public good, and in restricting the range of addressees, no one is excluded from an economically or socially relevant resource or opportunity. The restriction is thus logically exclusive, but not in a relevant context. In this case, then, the efficiency argument is sufficient to justify exclusion. The case might lie different, though, if a relevant good would be involved.

Scenario 3: Starting a photography group.

Setting: Since Sanaya is new to Italy, she does not yet have a local community. She wants to start a photography group to share skills and equipment, support each other in submitting works to art competitions, and recommend each other to important galleries. She asks through the chatbot: "Are there other semi-professional photographers interested in starting a photography group?". Sanaya asks people similar to her (e.g., other experienced photographers with good equipment), because she wants to cooperate on an advanced level rather than end up teaching amateurs.

Ethical reasoning and design action: Limit diversity, to support freedom of choice. Limiting diversity here is intuitive in order to find the right people for the envisioned group. This is an efficiency argument. However, even if this setting is again situated in the leisure sphere of self-government, a student group is an important space to create a community, learn skills, and pass opportunities to each other. Therefore, it is exclusive to limit diversity in this scenario and to make membership in the group indirectly dependent on owning the necessary equipment. This is an inclusion argument. Nonetheless, it is the student's free choice to meet with other experienced photographers rather than beginners or people without advanced equipment. The inclusion argument and the efficiency argument are in tension in this case, but since this case is related to the self-governance sphere, where freedom of choice is central, this last point ultimately trumps the inclusion argument.

Scenario 4: Paying for a service.

Setting: Daniel is studying German language and literature in Germany. He is also a passionate writer and has already won several essay competitions. Alongside his studies, he has written a novella that he plans to publish with a renowned University Press. Before submitting the manuscript, he wants to have someone proofread it. He plans to hire another student who seeks to become an editor. He inquires via the chatbot: "I am looking for a native German speaker to edit a 90-page literary text for €500. Interested?". He asks people who are similar to him (e.g., native German speakers who are studying in the field of literature), because he suspects that only native speakers are suitable for the editing task at hand.

Ethical reasoning and design action: Do not limit diversity, to support inclusion and to promote equality. In this case, the request relates to an economic and professional opportunity. Here, it is important that opportunities are at least potentially open to all people, regardless of their native background, in order to avoid contributing to prejudices and discrimination. By asking only German natives for help, Daniel excludes people with migration background from a good economic and professional opportunity. In this case, limiting diversity is ethically problematic due to equality reasons.

5.2 Resulting Ethical Norms

Our initial ethical analysis gave rise to four ethical norms, which we present below along with an intuitive description of their meaning:

Efficiency Norm: It is useful and it is not unethical to limit diversity if the request speaks to a certain demographic, or involves knowledge / skills tied to a specific social practice.

Safe-Space Norm: It is advisable to limit diversity if the request relates to the sensitive sphere, or is otherwise deemed to be sensitive due to certain relevant cultural considerations.

Freedom-of-Choice Norm: It is possibly problematic but not necessarily unethical to limit diversity if the request relates to the self-governance sphere, even if it might be exclusive.

Non-Discrimination Norm: It is unethical to limit diversity if the request relates to the public sphere, involving resources or decisions that should, in principle, be open for everyone.

6 STEP 3: IMPLEMENTATION OF ACTIONABLE DIVERSITY-CURATION POLICY

The third step calls for the technical implementation of the identified norms into an actionable policy. The fundamental task is to enable a machine to independently identify which norms apply in a given scenario (recognizing that multiple norms may be in tension) and to take specific actions in cases where intervention is appropriate. This task requires the designer to consider how to operationalize the identified ethical norms, how to resolve potential tensions between multiple applicable norms, and how the machine should intervene in a user's activities. Furthermore, the tension between interventions and the user's willingness to stick with the app (based on the user experience) is another important concern.

In the following, we make these considerations and discuss what an operationalization effort would entail, what technologies could be used, how various options could affect the end result, and, critically, the ethical ramifications of any choices that could be made.

6.1 Operationalizing the Ethical Norms

Each norm can be viewed as a pair of a condition and a conclusion, indicating, respectively, the (sufficient) circumstances under which the norm is applicable, and whether diversity can (not) be or should (not) be limited according to that norm. A key consideration in operationalizing the norms, then, is to determine the conditions of which norms hold in any given scenario. This ties to the choice of how the scenarios themselves are formally represented.

The obvious option is to represent a scenario in natural language using an open-ended vocabulary, as we have done in Section 3, the assumption being that, in the context of an automated system, the scenario is provided in textual form by the users themselves. One could then adopt Natural Language Processing (NLP) techniques [5] to identify the underlying syntactic structure and semantic content of a scenario of relevance when determining norm applicability.

An alternative would be for users to complete relevant fields in their profile to indicate their current situation, which might be cognitively more demanding for users, but could reduce the potential mistakes that would come from the automated parsing of the textual scenario. This approach leaves much to be desired. First, a user's profile, even if continuously populated with incoming data, is usually static in its basic structure. User questions, on the other hand, can vary greatly depending on mood and occasion, so the profile is insufficient as a context source for all questions. Second, the cognitively more demanding nature of the task suggests that the profile information might end up not being completely accurate (e.g., because of distracted users), raising the danger of suggesting a context quite distant from what is the case for a particular question.

Given the importance of question-specific context, we are led to suggest that textual information accompanying the question, and processed through NLP, might be the most effective way forward.

Settling on the representation of the scenario does not, yet, address the challenge of operationalizing norms, since we still need to decide how the conditions of the norms are themselves represented. Looking at the conditions of the ethical norms that resulted from the second step of our adopted strategy, we note that their main (although not the only) point of reference relates to the social spheres. A norm is applicable in a given scenario, if the question can be shown to refer to a situation within the norm's associated social sphere. This, in turn, would seem to suggest that some form of ontology could be developed, based on which different situations could be mapped to the different social spheres. By way of example, a question mentioning "stress" could be mapped to the class "mental conditions", and this hierarchically to the class "health", which can then be identified as a topic in the scope of the sensitive sphere.

Constructing such an ontology is not trivial. A small set of experts could presumably identify key classes in this ontology, perhaps using existing lexical resources as their basis [40, 52, 53], but it is unclear whether the ontology would ever be sufficiently complete to cope with future scenarios. Particularly at the lower levels of the ontology, the variability and the nuances of natural language make it practically impossible to be exhaustive. For example,

> "Following *a collaboration* / *an appointment* with a drug addiction specialist, I am looking to read more on the subject. Can you recommend relevant books?"

are two syntactically similar questions, but the former might more reasonably be classified as part of a professional activity and, thus, in the public sphere, whereas the latter might more reasonably be classified as part of the sensitive sphere. Beyond this challenge, it is also possible for a certain question to be classified in multiple spheres. If the latter question from above were to be extended with an offer to pay for the sought information, this could place the question also in the public sphere. This would entail that an ontology would not be strictly hierarchical, which increases the complexity of creating it and maintaining its coherence.

Instead of constructing an ontology, one could consider employing Machine Learning (ML) techniques, and train a model to map directly a scenario from its natural language (or its NLP-parsed) representation to the applicable norms. As is usually the case with ML, this would require large amounts of scenarios, each annotated by the applicable norms, and would produce an opaque mapping that critically relies on the quality of its training data. One option could be to outsource the creation and annotation of these scenarios to crowd-workers. However, the current situation with this form of work is that most common crowd-sourcing environments must be considered deeply exploitative. Therefore, in its current shape, this option is not suitable for an ethically-inspired project [24, 51, 65].

Nevertheless, crowd-sourcing might still be the most feasible way forward, as it enables the generation of a continuous stream of large and diverse training data to initiate a robust ML-based

process. It also leverages user experience as a ground truth, avoiding the pitfalls and limitations of a top-down, expert-driven approach. However, when using such an approach, it is essential to ensure that a best practice procedure is applied. This should not merely imply that benefits are shared appropriately among all stakeholders, but also the mandatory involvement of an ethics team to manage and monitor the procurement process. Last but not least, fair working conditions for the crowd-workers are an indisputable prerequisite.

The opacity of the resulting ML model could be alleviated through the use of neural-symbolic architectures [57]. Such architectures support the seamless integration of ML models with symbolically-represented knowledge, and would allow the ML model to focus on extracting from the natural language text those symbolic features that are relevant for determining norm applicability, while offloading the actual applicability check to the symbolic module.

6.2 Resolving Tensions Between Norms

Once we are able to determine which norms are applicable in a given scenario, we need to decide how to cope with tensions among norms, if multiple ones happen to apply concurrently. Tackling this issue is predicated on whether our identified norms have an inherent priority between them that holds immutable across scenarios (such as in the clear case of a fundamental value, such as equality, being in tension with an instrumental one, such as loyalty), or whether this priority is contextualized on nuances of the scenarios that are not captured by the conditions of the norms (such as when two fundamental or two instrumental values stand in tension).

If the norms are indeed inherently prioritized, then the balancing process can in principle be resolved upfront, through theoretical analysis performed by a small group of experts. Admittedly, this would be resource demanding, as at least one scenario would need to be identified for each pair (or possibly subset) of norms in tension. However, the number of scenarios is relatively manageable since we have a limited number of norms. Nonetheless, the challenge here could also be conceptual, raising the question of whether it is even possible to create meaningful scenarios from an expert position that sufficiently resemble relevant real-life situations.

If, then again, the priority between norms is contextualized on nuances of the scenarios, further theoretical analysis does not seem to be a feasible approach. One could consider the use of ML, but analogous, and in fact more resource-demanding, issues as those discussed in the operationalization of norms would come up.

In the case of contextualized norm priorities, one might need to accept that any balancing act needs to be resolved on a case-by-case basis. This task can be best assigned to the ultimate beneficiaries of our attempt to develop a "diversity-aware" system: the system's users. This could be realized by presenting a user with the set of all applicable norms in each particular scenario, and, in case of a tension produced by multiple applicable norms, let the user resolve that tension based on their intuitive understanding of their own situation. In a sense, the system still fulfills its purpose of curation by balancing protection and inclusion, as it raises awareness by alerting users that certain norms should be followed, but ultimately it is up to the users themselves to determine the path of the process.

In the context of our particular use-case, after users ask a question through the chatbot, they are not given a clear direction in terms of whether limiting diversity is ethically justified, but they are given norms supporting both sides of the "diversity dilemma", and are left to deliberate on which norm they will end up adopting. On the one hand, this might lead to even higher willingness to adjust their requested diversity according to the balancing of protection and inclusion, as the users are actively deliberating on the matter. On the other hand, the users might simply exhibit confirmation bias, where they overestimate the importance of an applicable norm that happens to agree with their already requested extent of diversity.

Ultimately, the question whether norm tensions can be resolved upfront by experts, or on a case-by-case basis by the users, or (more plausibly) through a hybrid approach, calls for an empirical investigation. In any case, the tension resolution mechanism could be understood through the prism of Formal Argumentation [13], and could be viewed as a cognitive assistant supporting the decision-making process of the users [28]. This would be especially pertinent if the norms end up increasing in number, with complex interactions between them that require some cognitive effort to be appreciated.

6.3 Determining Appropriate Interventions

Once the applicable norms (of highest priority) are identified in a given scenario, the final issue to be handled is how the conclusion of the norms is to be acted upon. What instruments are available to make the system's norm-driven policy actionable? The arena of options here is rather diverse, ranging from opt-in or opt-out textual or graphical messages, to nudges and pop-up notifications. Instead of presenting a list of concrete instruments to be used in the diversity curation process, we present five dimensions that have to be considered when making the relevant design choices.

How? One of the primary consideration is how forcefully one intervenes in cases of unjust exclusion. From an ethical perspective, it would be problematic to take control over the user's choices when interacting in a chatbot. However, nudges to the user can be milder or stronger, e.g., through the use of language and emotional appeal. On the more forceful side of things, the automated system might allow the user to make an initial selection (Do they want to send the request to someone similar or different?), and then, if ethically advisable, encourage the user to make a change in that selection.

Why? The intervention can be accompanied by varying degrees of justification. This could range from offering an explanation in unilateral support of the chosen intervention, to explaining all potential interventions. Note that information about the diversity curation process should be transparent to the user, as the system potentially affects the behavior of the user. To ensure transparency in a manner that is cognitively compatible with a given user, one could appeal to a machine learning process that learns by *being explained to / coached* by the user [38] and is explainable by design, rather than one that produces only post-hoc explanations [32, 47].

Who? This work has focused on the case of an automated system undertaking the monitoring role to decide whether an intervention is required. One could meaningfully consider, as a design choice, other candidates to take on this role. We have, for example, suggested earlier that part of this monitoring could be delegated to the user asking a question, or even to the community as a whole.

When? Analogously, this work has focused on the case of the intervention happening right after a user asks a question and selects if and how to restrict its dissemination, with the intervention targeting the user. One could also consider interventions at other points of the process, including prior to the user asking a question, or prior to the user being asked how to restrict its dissemination.

What? A final consideration relates to what form of intervention one is interested in. We have focused on the case of local interventions, associated with each question asked. Other options could be available, such as interventions at the time a user joins the platform. For example, information could be provided to increase *diversity literacy* when joining, so that users are aware of the social consequences of their decisions while interacting with the community.

7 LIMITATIONS AND LOOKING AHEAD

In this paper, we have adopted a value-centered scenario analysis to develop a strategy for curating diversity by balancing the protection and inclusion of users in an online social platform. This initiative is based on the recognition that sometimes we need to limit the diversity of communications in order to *(i)* prevent hate speech, silencing, discrimination, and *(ii)* promote mutual understanding, tolerance, and democracy in online communities. Following our three-step strategy, we have developed norms for the process of balancing inclusion and protection, and then considered different paths of how a machine could implement an actionable policy.

Our approach is limited by the fact that our proposed ethical intervention presents a technical fix. As discussed in Section 6, relying on ML and NLP raises ethical concerns about bias and conformity. This is where user responsibility comes in. Contemporary ethics emphasizes the importance of context for moral judgments, and STS research admonishes us not to fall into a simplistic binary thinking that pigeonholes the (social) world into an either / or logic and thus fails to recognize its chaotic, interactive, and fluid character [30]. So, while to some extent we need certain rules and a pre-structured environment that provides a baseline of protection, especially for the most vulnerable groups of users, we also need to focus on, acknowledge, and support user agency. Ultimately, users of an online community are the experts of their own lifeworlds.

Our analytic strategy seeks to accommodate these convictions as much as possible by recognizing and taking into account the different social spheres and situated actions of users. However, in what we can do on the computational side of things, we are still bound by the requirements of automation. Thus, it is hard to deny that the machine learning and reasoning tools used to develop the chatbot require us to abstract from context, and, instead, establish rules that are generalizable and formalizable, and to produce results that are unambiguous and binary enough to be translated into code.

Does this mean that the project of an ethics-driven, agency-centered, and context-sensitive "diversity by design" is a contradiction in terms? It definitely cannot be pursued without compromises and most probably not without flaws [4]. But since values are inscribed in technology anyway, and neutrality is an illusion [20], our approach makes the inscription of values with all its constraints as considerate, deliberate, and transparent as possible. Our formal take also supports the extension of inscribed values and ethical norms in an elaboration-tolerant manner, if and when the need might arise, potentially following future expert analyses, user-provided feedback [36, 37], or machine-learned knowledge [35, 39].

Looking ahead, this paper has not only presented a strategy, but also set a research agenda for the technical implementation of ethical norms and considerations to promote inclusion and protect diversity in online social platforms. Next steps include further iterations of our strategy to refine the scenarios and the ethical norms, and to formally codify them. We also expect that additional scenarios will need to be considered as the work progresses, due to the numerous value tensions that arise from a combination of values. Future work will need to address the specific socio-technical, aesthetic, and political tools that can be used to tackle the challenge of balancing inclusion and protection in social platforms.

ACKNOWLEDGMENTS

This work was supported by funding from the EU's Horizon 2020 Research and Innovation Programme under grant agreements no. 739578 and no. 823783, and from the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation, and Digital Policy. The authors would like to thank Martel Innovate for the graphic art, and the anonymous reviewers for valuable feedback.

REFERENCES

[1] Himan Abdollahpouri. 2019. Popularity Bias in Ranking and Recommendation. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Vincent Conitzer, Gillian Hadfield, and Shannon Vallor (Eds.). ACM, New York, NY, USA, 529–530. https://doi.org/10.1145/3306618.3314309

[2] Regine Bendl, Edeltraud Hanappi-Egger, and Roswitha Hofmann. 2012. *Diversität und Diversitätsmanagement* (1. aufl. ed.). Utb-studi-e-book, Vol. 3519. UTB GmbH, Stuttgart.

[3] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code* (1. edition ed.). Polity.

[4] Wiebe Bijker. 2017. Constructing Worlds: Reflections on Science, Technology and Democracy (and a Plea for Bold Modesty). *Engaging Science, Technology, and Society* 3 (2017), 315–331. https://doi.org/10.17351/ests2017.170

[5] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly Media, Inc.

[6] Axel Bruns. 2019. Filter Bubble. *Internet Policy Review* 8, 4 (2019). https://doi.org/10.14763/2019.4.1426

[7] Manuel Castells. 2009. *The Rise of the Network Society: The Information Age: Economy, Society, and Culture Volume I* (2. edition ed.). Wiley-Blackwell.

[8] Nicole Chi, Emma Lurie, and Deirdre K. Mulligan. 2021. Reconfiguring Diversity and Inclusion for AI Ethics. (2021). https://doi.org/10.1145/11952.107 arXiv:2105.02407

[9] Sasha Costanza-Chock. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need.* The MIT Press.

[10] Kimberlé Crenshaw. 1988. Race, Reform, and Retrenchment: Transformation and Legitimation in Antidiscrimination Law. *Harvard Law Review* 101, 7 (1988), 1331–1387.

[11] Jessie Daniels. 2009. *Cyber Racism: White Supremacy Online and the New Attack on Civil Rights.* Rowman & Littlefield Publishers, Lanham and Boulder and New York and Toronto and Plymouth, UK.

[12] Marco Dean. 2019. Scenario Planning: A Literature Review. *https://www.doi.org/10.13140/RG.2.2.12629.24802* (2019).

[13] Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence* 77, 2 (Sept. 1995), 321–357. https://doi.org/10.1016/0004-3702(94)00041-X

[14] Virginia Eubanks. 2018. *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor* (illustrated edition ed.). St Martin's Press.

[15] European Commission. 1999. The EU Values. https://ec.europa.eu/component-library/eu/about/eu-values/

[16] Berenice Fisher and Joan Tronto. 1990. Toward a Feminist Theory of Caring. *Circles of Care: Work and Identity in Women's Lives* (1990), 35–62.

[17] Luciano Floridi. 2010. *The Cambridge Handbook of Information and Computer Ethics.* Cambridge University Press.

[18] Rainer Forst. 2013. *Toleration in Conflict: Past and Present.* Cambridge University Press. https://doi.org/10.1017/CBO9781139051200

[19] Batya Friedman and David Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination.* The MIT Press, Cambridge, MA.

[20] Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Transactions on Information Systems* 14, 3 (1996), 330–347. https://doi.org/10.1145/230538.230561

[21] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (illustrated edition ed.). Yale University Press.

[22] Tarleton Gillespie. 2020. Content Moderation, AI, and the Question of Scale. *Big Data and Society* 7, 2 (July 2020). https://doi.org/10.1177/2053951720943234

[23] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. *Big Data and Society* 7, 1 (2020), 1–15.

[24] Mary L. Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Houghton Mifflin Harcourt, Boston.

[25] Natali Helberger. 2019. On the Democratic Role of News Recommenders. *Digital Journalism* 7, 8 (2019), 993–1012. https://doi.org/10.1080/21670811.2019.1623700

[26] Paula Helm. 2017. What Can Self-Organised Group Therapy Teach Us About Anonymity? *Ephemera. Theory and Politics in Organization* 17/2 (2017), 327–350.

[27] Paula Helm. 2018. Treating Sensitive Topics Online: A Privacy Dilemma. *Ethics and Information Technology* 20, 4 (Dec. 2018), 303–313. https://doi.org/10.1007/s10676-018-9482-4

[28] Antonis C. Kakas and Loizos Michael. 2016. Cognitive Systems: Argument and Cognition. *IEEE Intelligent Informatics Bulletin* 17, 1 (2016), 14–20.

[29] Cory Knobel and Geoffrey C. Bowker. 2011. Values in Design. *Commun. ACM* 54, 7 (2011), 26–28. https://doi.org/10.1145/1965724.1965735

[30] John Law. 2007. *After Method: Mess in Social Science Research.* Routledge, London.

[31] Karen Lumsden and Emily Harmer. 2019. *Online Othering: Exploring Digital Violence and Discrimination on the Web.* Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-12633-9

[32] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems.* 4768–4777.

[33] Catharine A. MacKinnon. 1989. *Toward a Feminist Theory of the State.* Harvard University Press.

[34] Ishani Maitri. 2009. Silencing Speech. *Canadian Journal of Philosophy* 39, 2 (2009), 309–338. https://doi.org/10.1353/cjp.0.0050

[35] Loizos Michael. 2016. Cognitive Reasoning and Learning Mechanisms. In *Proceedings of the 4th International BICA Workshop on Artificial Intelligence and Cognition (CEUR Workshop Proceedings 1895).* New York City, New York, U.S.A., 2–23.

[36] Loizos Michael. 2019. Machine Coaching. In *Proceedings of the IJCAI Workshop on Explainable Artificial Intelligence.* Macao SAR, P.R. China, 80–86.

[37] Loizos Michael. 2020. Machine Ethics through Machine Coaching. In *Proceedings of the 2nd Workshop in Implementing Machine Ethics.*

[38] Loizos Michael. 2021. Explainability and the Fourth AI Revolution. *CoRR* abs/2111.06773 (2021). arXiv:2111.06773

[39] Loizos Michael and Leslie G. Valiant. 2008. A First Experimental Demonstration of Massive Knowledge Infusion. In *Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning*, Gerhard Brewka and Jérôme Lang (Eds.). AAAI Press, Sydney, Australia, 378–389.

[40] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[41] Linda Monsees. 2021. Information Disorder, Fake News and the Future of Democracy. *Globalizations* (2021), 1–16. https://doi.org/10.1080/14747731.2021.1927470

[42] Lisa P. Nathan, Predrag V. Klasnja, and Batya Friedman. 2007. Value Scenarios: A Technique for Envisioning Systemic Effects of New Technologies. In *Proceedings of the CHI 2005 Extended Abstracts on Human Factors in Computing Systems.* ACM, 2585–2590. https://doi.org/10.1145/1240866.1241046

[43] Cathy O'Neil. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (1st edition ed.). Penguin Books UK.

[44] Eli Pariser. 2012. *The Filter Bubble: What the Internet is Hiding From You.* Penguin Books, London.

[45] Maya Ranganathan. 2019. Re-Scripting the Nation in 'Post Truth' Era: The Indian Story. *Asian Ethnicity* (2019), 1–15.

[46] Andreas Reckwitz. 2002. Toward a Theory of Social Practices. *European Journal of Social Theory* 5, 2 (2002), 243–263. https://doi.org/10.1177/1368431022225432

[47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 1135–1144.

[48] Sarah T. Roberts. 2019. *Behind the Screen.* Yale University Press.

[49] Laura Schelenz, Matteo Busso, Ivano Bison, Amalia de Götzen, Daniel Gatica-Perez, Fausto Giunchiglia, Lakmal Meegahapola, and Salvador Ruiz-Correa. 2021. The Theory, Practice, and Ethical Challenges of Designing a Diversity-Aware Platform for Social Relations. In *Proceedings of the 4th AAAI/ACM Conference on Artificial Intelligence, Ethics and Society.* https://doi.org/10.1145/3461702.3462595

[50] Paul J. H. Schoemaker. 1995. Scenario Planning: A Tool for Strategic Thinking. *HMIT Sloan Management Review* 36, 2 (1995).

[51] M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar. 2018. Responsible Research with Crowds: Pay Crowdworkers at Least Minimum Wage. *Commun. ACM* 61, 3 (2018), 39–41. https://doi.org/10.1145/3180492

[52] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 4444–4451.

[53] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. YAGO: A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web* (Banff, Alberta, Canada). New York, NY, USA, 697–706.

[54] Charles Taylor. 2004. *Modern Social Imaginaries.* Duke University Press.

[55] Charles Taylor. 2007. *A Secular Age.* Harvard University Press.

[56] Alexander Tischbirek. 2017. *Die Verhältnismäßigkeitsprüfung* (1 ed.). Studien und Beiträge zum Öffentlichen Recht, Vol. 35. Mohr Siebeck. X, 239 pages.

[57] Efthymia Tsamoura, Timothy M. Hospedales, and Loizos Michael. 2021. Neural-Symbolic Integration: A Compositional Perspective. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence.* AAAI Press, 5051–5060.

[58] Anna Lowenhaupt Tsing. 2004. *Friction: An Ethnography of Global Connection.* Princeton University Press.

[59] UNESCO. 2005. The Convention on the Protection and Promotion of the Diversity of Cultural Expressions. *https://en.unesco.org/creativity/convention* (2005).

[60] Tommaso Venturini. 2019. From Fake to Junk News: The Data Politics of Online Virality. In *Data Politics.* Routledge, 123–144.

[61] Steven Vertovec. 2012. "Diversity" and the Social Imaginary. *European Journal of Sociology* 53, 3 (2012), 287–312. https://doi.org/10.1017/S000397561200015X

[62] Jeremy Waldron. 2012. *The Harm in Hate Speech.* Harvard University Press, Cambridge, MA.

[63] Michael Walzer. 1984. *Spheres of Justice: A Defense of Pluralism and Equality.* Basic Books, New York.

[64] WeNet: The Internet of Us. 2022. Project Page: https://www.internetofus.eu/.

[65] Jamie Woodcock and Mark Graham. 2020. *The Gig Economy: A Critical Introduction* (1 ed.). Polity.

[66] Michael J. Zimmerman and Ben Bradley. 2019. Intrinsic vs. Extrinsic Value. In *The Stanford Encyclopedia of Philosophy* (Spring 2019 ed.), Edward N. Zalta (Ed.). Stanford University.