

# Diversity by Design: Balancing Protection and Inclusion in Social Networks<sup>\*</sup>

Paula Helm<sup>1</sup>, Loizos Michael<sup>2,3</sup>, and Laura Schelenz<sup>1</sup>

<sup>1</sup> International Center for Ethics in the Sciences and Humanities

University of Tübingen, Tübingen, Germany

{paula.helm, laura.schelenz}@uni-tuebingen.de

<sup>2</sup> Open University of Cyprus, Nicosia Cyprus

<sup>3</sup> CYENS Center of Excellence, Nicosia, Cyprus

loizos@ouc.ac.cy

**Abstract.** The unreflected promotion of diversity as a value in social interactions — including technology-mediated ones — risks emphasizing the benefits of inclusion at the cost of not recognizing the potential harm from failing to protect stigmatized or marginalized individuals. Adopting the stance that technology is not value-neutral, we attempt to answer the question of how technology-mediated social platforms could accommodate *diversity by design*, by balancing the often competing values of protection and inclusion. This short paper presents our research agenda as well as initial analysis and outcomes. Building on approaches from scenario planning and the methodology of Value Sensitive Design, we identify ethical principles and arguments on how to curate diversity, which we seek to operationalize through formal argumentation.

## 1 Diversity in Social Media, Too Much of a Good Thing?

While modern digital technologies hold an enormous potential to bring together people with different skills, experiences, opinions, backgrounds, competencies, and resources, helping transcend geographic boundaries in connecting people, the unchecked adoption of diversity can end up becoming a hindrance to meaningful online engagement. Unsurprisingly, then, designing systems that strike a fine balance between the protection and the inclusion of individuals is a challenging task, with existing social media and networking platforms drawing criticism both for *(i)* failing to take advantage of the diversity of their users, and *(ii)* failing to curtail the negative sides of diversity, such as abusive behavior [22].

In order to tackle this criticism from an interdisciplinary Ethics and Computer Science perspective, we explore the socio-technical possibilities of intervening in a diversity-aware chatbot to empower users. Our guiding research question is: *Can we build algorithms to support meaningful user interactions on social networking platforms in a way that considers the benefits and risks of diversity?*

We attempt to answer this question in the context of the EU-funded project “WeNet – The Internet of Us” (<https://www.internetofus.eu>), where the goal

---

<sup>\*</sup> All authors contributed equally to this paper.

is to facilitate machine-mediated diversity-aware social interactions. As part of the project’s ongoing pilot studies, a chatbot application has been developed that supports users — university students, in the current pilots — across geographic, social, and cultural backgrounds in seeking help from the “crowd” of other platform users through written messages, with the expectation that the platform will be considerate of sensitivities and power asymmetries.

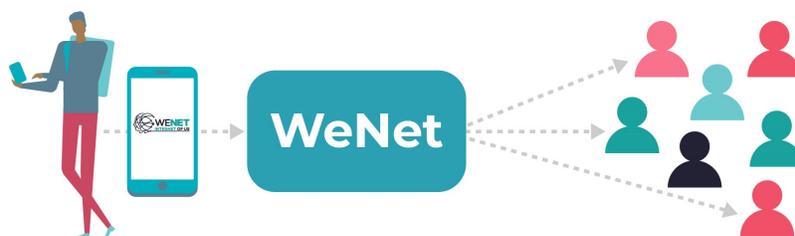
The design of the platform and the chatbot follows the “Values in Design” research program [7, 13], which opposes the widespread attitude of technology being value-neutral. Accordingly, the project seeks to carefully identify the values to be inscribed in the developed technologies, and what implications this process will have when materialized in the real world. In this spirit, this work considers the use-case of a chatbot to reflect on and discuss our posed question.

## 2 Ethical Diversity Curation when “Asking for Help”

In society, diversity is often invoked to create associations with pluralism, tolerance, or simply “good” policy. Diversity has become popular as a marketing strategy and been adopted in the rhetoric of big tech companies [2]. Yet, diversity is more complicated than usually insinuated in public discourses. From an ethical perspective, it is important to critically review diversity concepts and pay attention to the negative implications of unfiltered diversity. While diversity in technology-mediated spaces does not necessarily differ from diversity in society, the mediation raises questions about the responsibility of technology designers and providers to curate diversity. This is similar to political actors’ responsibility to, e.g., protect minorities in liberal democracies [27]. Even though popular technology may be designed by a number of powerful private actors, the quality of exchange in platforms today resembles the one in public spaces. It is therefore ethically relevant to regulate online and communication spaces to prevent harm to users and allow for democratic opportunities to negotiate these spaces [3].

As per the WeNet project, we understand diversity as a descriptive and normative concept. In the first direction, diversity is a concept that helps us define differences between users and ultimately classify elements of their being (e.g., skills and practices) that can complement other users’ characteristics to form a moment of mutual support. In the second direction, we treat diversity as a value and differentiate between ideas of diversity as intrinsic value (diversity is important for itself) and diversity as instrumental value (diversity helps achieve other values such as inclusion or productivity) [30]. This differentiation is relevant for the ethical intervention in the diversity-aware WeNet chatbot and platform. There, we build on our understanding of diversity as an instrumental value that supports inclusion, tolerance, and justice (cf. Section 3). However, diversity also needs to be weighted against other demands such as protection and efficiency.

Furthermore, diversity can be seen as a strategy, which nicely aligns with our understanding of diversity as an instrumental value. In recommender systems, for example, diversity has been established as the practice of diversifying the item sets recommended to users in order to increase user satisfaction [20], or



**Fig. 1.** Visual representation of the WeNet technology: A user can ask any question to a diverse community of peers, through a machine-powered social platform.

foster democratic principles [26]. Helberger discusses diversity in the context of media and news recommendations, and argues that diversity is a goal but also a design strategy to enable different choices and hence increase autonomy of users [10, 11]. In that context, the term “Diversity by Design” is used to refer to “the idea that it is possible to create an architecture or service that helps people to make diverse choices” [10]. We adopt the term “Diversity by Design” in the title of this paper and expand further the idea to include not only diverse choices, but also safe, autonomous, and beneficial interaction in our envisioned chatbot.

The chatbot “Ask for Help” is designed as an instant messaging application within Telegram, running over the WeNet diversity-aware platform (cf. Figure 1). The chatbot allows users to send questions (or “requests”) to the community of registered users of the chatbot. When a user poses a question to the community via the command “/question”, they are requested to specify whether they want to ask users who are “similar” to them or who are “different”. While this request initially served research purposes to define a better understanding of users’ needs for diversity, it also allows users to specify the scope of respondents. The queries by users are analyzed algorithmically on the basis of previously collected user profiles in an attempt to determine the most appropriate users to respond.

In this step, the system relies on the descriptive conceptualization of diversity (see the discussion above); data has been collected through self-report mechanisms as well as mobile sensor and geolocation tracking to determine a user’s set of skills and practices. This sociological concept of diversity goes beyond classic demographic approaches to building user profiles and thus may enrich the user’s experience with the chatbot’s matching to other users; cf. [22]. Ideally, the chatbot thus takes into account aspects of both fit (diversity of skills and practices; i.e., who has the skills to provide an answer to my request?) and individual preferences for diversity (i.e., do I need users with similar experiences to provide for a safe space, or can I benefit from expertise that complements my own?).

The ethical intervention in the chatbot was developed after a first round of experiments revealed that students were asking sensitive questions about mental health, on the one hand, but also covering highly controversial topics around the

COVID-19 pandemic or religion, on the other hand. To mitigate potential harm to users from, e.g., hate speech and abuse, or insensitive responses that provoke a deterioration of vulnerable users’ mental health, the ethics and design team decided to introduce extra algorithmic measures and design instruments, such as nudges and report mechanisms, to offer users more opportunities for protection.

### 3 Methodology: Value-Sensitive Scenario Analysis

In our research, we combine approaches from the field of scenario planning [5, 23], with the methodology of Value Sensitive Design (VSD) [7]. Through these techniques and methods, we seek, iteratively, and in an agile manner, to:

- (1) conceptualize the relevant and desirable values, and identify the potential tensions between them, for ethically accommodating diversity in the chatbot;
- (2) empirically determine, primarily through scenario thinking, how the relevant values are weighted against each other in case of incompatibilities;
- (3) technically implement the outcomes of the earlier phases by operationalizing the identified values and their weights in a formal representation language.

As part of phase (1), we approach diversity as a normative value specifying “what ought to be”, and also as an instrumental value to be leveraged to achieve other important values that empower users and benefit the society: promoting numerous opinions and political orientations, achieving equity and social justice [4], expanding experiences and knowledge, fostering tolerance and mutual understanding [25], and advancing business goals such as solving complex tasks by bringing together diverse team members in productive ways [24].

Juxtaposed with these benefits of diversity are several risks: perceiving social signals as incoherent noise, harming or silencing minorities or marginalized groups through hate speech [16], and being non-conducive to meaningful connectivity by excluding safe spaces for sensitive issues such as health [12]. Intriguingly, sacrificing a certain level of diversity (e.g., in expressed opinions) to avoid the risks above, ends up promoting and protecting diversity (e.g., by supporting minorities to express freely), which is key to equality in liberal societies [27].

As part of phase (2), we empirically determine and carefully assess the contexts in which diversity unfolds, and whether its limitation should be promoted, avoided, or dismissed as unethical. Inspired by real user profiles and requests from the WeNet project’s first pilot, we draw on methods of scenario analysis [29] to develop four scenarios and supplement them with two fictional ones to explore the fringe cases of potential misuse. Each scenario (cf. Figure 2) includes a user’s background and request, and whether they chose to limit the respondents.

Furthermore, drawing on the common distinction between instrumental and fundamental values [30], with the latter being intrinsically valuable and, therefore, weighing more heavily than the former, we identified various values touched upon in the scenarios. The values are: (*instrumental values*) inclusion, tolerance, freedom of choice, efficiency; (*fundamental values*) autonomy, well-being, health, dignity, justice. Loosely inspired by Walzer’s analysis in “Spheres of Justice” [28],



**Fig. 2.** An example scenario which stands for requests made through the “Ask for Help” chatbot that are sensitive and may require a safe space. In this particular scenario, we determine that limiting diversity, i.e., the scope of possible respondents to the user’s question, is legitimate. Of course, this means that the system discriminates against some other chatbot users by excluding them from being considered as potential respondents, but the need for protection of the user asking the question in this particular case outweighs the risks of exclusion of others from consideration as respondents.

we also take into account the contextual dependency of value judgments, especially when it comes to balancing conflicting values at tension with each other.

Ideal-typically, we distinguish three spheres to which the different scenarios relate. In analyzing the relative relevance of the sometimes competing values, we take the different requirements (protection and inclusion) corresponding to the different spheres into account. We are aware that the boundaries of the spheres we present in this framework are fluid in reality and in fact often difficult to clearly delineate, but for our analysis they provide the necessary analytical basis to make a reasoned decision in ambivalent cases. The spheres are: (*maximum freedom sphere*) leisure / sports / art; (*shared resources sphere*) economy, politics, education; (*protection / sensitive sphere*) religion, health / medicine, psychology.

Finally, in addition to the various values and the social spheres, we consider the economic-historical-cultural situatedness from which users make inquiries and from which these inquiries are answered [9]. The six scenarios are analyzed along these three considered dimensions (i.e., values, social spheres, and situatedness) to determine how diversity should be curated in each case and why.

#### 4 Operationalizing Ethics through Formal Argumentation

When the first iteration of phases (1)–(2) of our value-sensitive scenario analysis is completed, we can proceed, as part of phase (3), to operationalize our findings.

We approach this operationalization by utilizing formal argumentation [6], turning the identified conditions under which diversity should be curated into arguments, and the relative weights that result from the considered values into an attacking relation between the arguments. Applying standard formal argumentation semantics helps decide how competing values that might be promoted in some context are to be handled when determining how to curate diversity.

The preliminary arguments stemming from our initial ethical analysis are:

**Efficiency Argument:** 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.

**Protection Argument:** It is imperative to limit diversity if the request relates to the protection / sensitive sphere or is sensitive due to a cultural / social context.

**Inclusion Argument:** It is unethical to limit diversity if the request relates to the shared resources sphere.

**Freedom-of-Choice Argument:** It is possibly problematic but it is not strictly unethical to limit diversity if the request relates to the maximum freedom sphere even if it might be exclusive.

**No-Harm Argument:** The question of curating diversity is irrelevant if the request is violating a fundamental ethical principle at any rate.

Once a determination is made on which argument prevails and how diversity is to be curated, various instruments can be used to implement the associated curation action: nudging users to revise their request, ensuring that users have a sufficiently-rich set of options on limiting the respondents to their requests, allowing users to block received requests deemed intrusive or offensive, or even offering a mechanism to launch complaints or report unethical requests.

Our ongoing work as part of phase (3) involves the use of the Web-STAR system [21] as a substrate to formally encode the aforementioned scenarios and arguments. Our working hypothesis is that the use of an argumentation-based representation and reasoning system will help us make concrete the relevant aspects of contexts and how those relate to the premises of the arguments, and to support, if needed, a further iteration of the value-sensitive scenario analysis.

Ultimately, we envision this line of work to produce a knowledge-based system that automatically determines how to ethically curate diversity in social networking contexts such as those of our chatbot use-case. The system will receive information about the user making a request and the request itself, expressed in some structured form or a controlled natural language [14], and will provide transparency on the decisions made. Through Machine Coaching [17, 18] the system will facilitate its in-situ ethical elaboration and fine-tuning [19], while accommodating the encoding of suggestions on which of the instruments could be more appropriately used to implement each diversity curation action.

## 5 Limits of Ethical Intervention and Looking Ahead

While our research agenda centers on the need to ethically align diversity in social networking technology, we recognize the limits of ethical intervention, especially when this intervention is expressed and encoded algorithmically. Contemporary Ethics goes, in fact, to great lengths to emphasize the importance of context for moral judgments and to admonish us not to fall into a kind of simplistic binary thinking that pigeonholes the (social) world into an “either/or” logic, thus failing to recognize its indeed messy, intersectional, and fluid nature [15].

While our framework of analysis attempts to accommodate these convictions as much as possible by taking into account different social spheres and the economic-historical-cultural situatedness of users, we are still bound by the demands of automation. Machine learning and reasoning tools used to develop the chatbot require us to *(i)* establish rules that are generalizable and formalizable, and *(ii)* enable outcomes that are clear enough to be translated into code.

So, is the “Ethics by Design” project not, thus, a contradiction in terms? It is definitely not possible to pursue this project without compromises, and probably not without mistakes [1]. However, since values are inscribed in technology anyway and since neutrality is an illusion [8], our approach makes the inscription of values with all its constraints as considerate, conscious, and transparent as possible. The use of formal argumentation further supports the extension of the inscribed values in an elaboration tolerant manner, if and when the need arises.

Looking ahead, this paper has set a research agenda for the technical implementation of ethical arguments and considerations to promote diversity in social media and networking platforms. The next steps include the refinement of scenarios and ethical arguments as well as the formal codification of the arguments. Additional next steps will also need to address the specific technical, aesthetic, social, and policy instruments that can be used to curate diversity in terms of both balancing and promoting inclusion and protection in social networks.

As we move forward, we expect that the work will likely require considering additional scenarios due to the number of value tensions arising from a combination of values. While in this paper we have limited ourselves to a selection of scenarios that are related to the issue of weighing protection against inclusion, other important value tensions include, for example, freedom and justice, or privacy and transparency. We will further focus on other issues derived from empirical data from the anticipated second pilot of the WeNet project.

*Acknowledgements:* 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 of the KI 2021 Workshop on AI and Ethics for their valuable feedback.

## References

1. Wiebe Bijker. Constructing Worlds: Reflections on Science, Technology and Democracy (and a Plea for Bold Modesty). *Engaging Science, Technology, and Society*, 3:315–331, 2017.
2. Nicole Chi, Emma Lurie, and Deirdre K. Mulligan. Reconfiguring Diversity and Inclusion for AI Ethics. <https://arxiv.org/pdf/2105.02407>, 2021.
3. Kate Crawford. Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics. *Science, Technology, & Human Values*, 41(1):77–92, 2016.
4. Kimberlé Crenshaw. Race, Reform, and Retrenchment: Transformation and Legitimation in Antidiscrimination Law. *Harvard Law Review*, 101(7):1331–1387, 1988.
5. Marco Dean. Scenario Planning: A Literature Review. <https://www.doi.org/10.13140/RG.2.2.12629.24802>, 2019.
6. Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77(2):321–357, September 1995.
7. Batya Friedman and David Hendry. *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press, Cambridge, MA, 2019.
8. Batya Friedman and Helen Nissenbaum. Bias in Computer Systems. *ACM Transactions on Information Systems*, 14(3):330–347, 1996.
9. Donna Haraway. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3):575, 1988.
10. Natali Helberger. Diversity by Design. *Journal of Information Policy*, 1:441–469, 2011.
11. Natali Helberger. On the Democratic Role of News Recommenders. *Digital Journalism*, 7(8):993–1012, 2019.
12. Paula Helm. Treating Sensitive Topics Online: A Privacy Dilemma. *Ethics and Information Technology*, 2:303–313, 2018.
13. Cory Knobel and Geoffrey C. Bowker. Values in Design. *Communications of the ACM*, 54(7):26–28, 2011.
14. Tobias Kuhn. A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, 40(1):121–170, March 2014.
15. John Law. *After Method: Mess in Social Science Research*. International Library of Sociology. Routledge, London, 2007.
16. Ishani Maitri. Silencing Speech. *Canadian Journal of Philosophy*, 39(2):309–338, 2009.
17. Loizos Michael. The Advice Taker 2.0. In *Proceedings of the 13th International Symposium on Commonsense Reasoning*, London, U.K., 2017.
18. Loizos Michael. Machine Coaching. In *Proceedings of the IJCAI Workshop on Explainable Artificial Intelligence*, pages 80–86, Macao SAR, P.R. China, 2019.
19. Loizos Michael. Machine Ethics through Machine Coaching. In *Proceedings of the 2nd Workshop in Implementing Machine Ethics*, 2020.
20. Seiki Miyamoto, Takumi Zamami, and Hayato Yamana. Improving Recommendation Diversity Across Users by Reducing Frequently Recommended Items. In *Proceedings of the 2018 IEEE International Conference on Big Data*, pages 5392–5394. IEEE, 2018.
21. Christos Rodosthenous and Loizos Michael. Web-STAR: A Visual Web-based IDE for a Story Comprehension System. *Theory and Practice of Logic Programming*, 19(2):317–359, 2019.

22. Laura Schelenz, Matteo Busso, Ivano Bison, Amalia de Götzen, Daniel Gatica-Perez, Fausto Giunchiglia, Lakmal Meegahapola, and Salvador Ruiz-Correa. 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*, 2021.
23. Paul J. H. Schoemaker. Scenario Planning: A Tool for Strategic Thinking. *MIT Sloan Management Review*, 36(2), 1995.
24. Alina Tugend. The Commonality of A.I. and Diversity. *New York Times*, 6 November 2018.
25. UNESCO. The Convention on the Protection and Promotion of the Diversity of Cultural Expressions. <https://en.unesco.org/creativity/convention>, 2005.
26. Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. Recommenders with a Mission. In Falk Scholer, Paul Thomas, David Elsweiler, Hideo Joho, Noriko Kando, and Catherine Smith, editors, *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 173–183, New York, NY, USA, 2021. ACM.
27. Jeremy Waldron. *The Harm in Hate Speech*. The Oliver Wendell Holmes Lectures, 2009. Harvard University Press, Cambridge, MA, 2012.
28. Michael Walzer. *Spheres of Justice: A Defense of Pluralism and Equality*. Basic Books, New York, 1984.
29. Emily York and Shannon N. Conley. Creative Anticipatory Ethical Reasoning with Scenario Analysis and Design Fiction. *Science and Engineering Ethics*, 26(6):2985–3016, 2020.
30. Michael J. Zimmerman and Ben Bradley. Intrinsic vs. Extrinsic Value. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2019 edition, 2019.