


Transforming Toxic Debates towards European Futures: Technological Disruption, Societal Fragmentation, and Enlightenment 2.0

Mehmet Ali Üzelgün

University Institute of Lisbon and Nova University
Lisbon, Portugal

 0000-0003-4426-9055

Iliana Giannouli

National and Kapodistrian
University of Athens, Greece

 0009-0000-2634-351X

Ioanna Archontaki

National and Kapodistrian
University of Athens, Greece

 0000-0001-6712-0531

Klára Odstrčilová

Charles University, Czech Republic

 0000-0002-8860-5869

Barbara Thomass

Ruhr University Bochum, Germany

Cláudia Álvares

University Institute of Lisbon, Portugal

 0000-0002-2882-5114

Abstract: Online toxicity refers to a spectrum of problematic communicative phenomena that unfold in various ways on social media platforms. Most of the current efforts to contain it focus on computational techniques to detect online toxicity and build a regulatory architecture. In this paper, we highlight the importance of focusing on the social phenomena of toxicity, and particularly, exploring the public understanding and future imaginaries of toxic debates. To explore how users construe online toxicity and envisage the future of online discussions, we examine 41 scenarios produced by European experts from the field of technology and culture. Through a content analysis informed by a narrative approach and insights from futures studies, we identify three myths that characterize the future scenarios: technological disruption, societal fragmentation, and digital Enlightenment. After a discussion of their relations, we conclude by stressing the importance of platform transparency and user empowerment.

Keywords: Toxic debates, topic-driven toxicity, future scenarios, algorithmic disruption, regulation of social media content

INTRODUCTION

An article published on the WIRED Magazine website on 8th January 2024 argues that we are entering “a digital dark age”, as online trust collapses due to several transformations in the online landscape (Neff, 2024). While Neff (2024) mentions developments around generative AI and associated issues, the article mainly highlights the lack of transparency of the social media platforms. Neff (2024) argues that their increasingly restrictive data access policies hamper the independent initiatives that monitor misinformation, harmful content, and deep fake propaganda, elevating the risks of online manipulation and polarization. Moreover, these phenomena are seen to be connected to the “attention economy” (Williams, 2018), in which every user action is measured, processed, and aggregated to become part of some commercial strategy. Perhaps the first lesson social media algorithms have learnt, in this economy, is that the more provocative a message is to a user, the greater the chances of capturing their attention. Research has shown that news stories conveying emotions of anger and surprise are shared through social media with greater frequency and speed (Fan & Gordon, 2014; Ferrara & Yang, 2015). The same goes for populist messages that provoke anger (Hameleers et al., 2017), and emotional posts in general (Stieglitz & Dang-Xuan, 2013). Findings also suggest that platform algorithms enhance emotional, partisan and polarizing content, particularly tweets expressing anger and animosity towards out-groups (Milli et al., 2023).

Platformization, the creation of hyper-interactive digital ecosystems that connect people across geographic boundaries, has been widely embraced due to its potential to foster democratic discourse and deliberative democratic ideals (Mendis, 2021). However, social media platforms have also brought about an escalation of phenomena grouped under the label “online toxicity” (Pascual-Ferrá et al., 2021; Rossini, 2019). The label is sometimes used synonymously with hate speech, involving “intentional statements or messages with discriminatory content” (Petlyuchenko et al., 2021, p. 114). At other times it may refer to all sorts of harmful content including extremism, bullying, trolling, harassment, physical threats, and online stalking (Patel et al., 2021), “obscenity, insults, and identity-based hate” (Adams et al., 2017, p. 1), and “rude language, harsh criticisms, anger, hatred, even threats” (Suler, 2004, p. 321). There is no doubt that online hate and toxicity have serious impacts on the willingness to participate in public debate, formation of personal and public opinion, and people’s interpretation of polarization around issues of common concern (Anderson et al., 2018).

The documentation behind Google’s Perspective¹ defines toxicity as “rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion” (Jigsaw, 2024). Accordingly, while “hate speech” or “abuse” refer to “specific categories of language that violate certain terms of service or laws”, the term toxicity is preferred in that it “refers to a broad category of language that is subject to individual interpretation” (Jigsaw, 2024). In this paper we examine how experts view online toxicity and envisage the future of online discussions. To study how users construe toxicity and envisage networked communication, we explore 41 future scenarios produced in four scenario-building workshops and a scenario-writing exercise within the frame of the EUMEPLAT Horizon 2020 project (see Table 1). Before we report about our study of future scenarios for “toxic debates”, we will provide an overview of the efforts to moderate and regulate social media content. Next, we will highlight the limitations of these efforts and argue that there are aspects of toxicity to be addressed at the level of the whole debate, without being stripped of its public-political content, and, more broadly, as a matter of online culture. We aim to study toxicity as a socio-communicative issue, captured by the notion of “toxic debates”, rather than as an interpersonal issue of psychological harm. Our content analysis of scenarios draws on a narrative perspective, congruent with futures studies (Hänninen et al., 2022; Inayatullah, 2008). We discuss three myths as the output of our content analysis: (i) ‘technological disruption’ refers to the impact of algorithms on platformed interactions, (ii) the primary impact of ‘societal fragmentation’, and (iii) ‘enlightenment 2.0’ that refers to the efforts to address and alleviate that impact. In our conclusion, we stress the importance of platform transparency and user empowerment.

MODERATION OF SOCIAL MEDIA CONTENT

Two interconnected trends can be identified in the treatment of the diverse phenomena grouped under the notion of online toxicity. The first is to treat toxicity as an interpersonal phenomenon with a source (offender) and a target (victim). The second is to treat it as a matter of rude, coarse, or “abusive language” (Nobata et al., 2016; Waseem et al., 2017) or as a verbal act, i.e., verbal aggression (Guberman et al., 2016; Kumar et al., 2018). The act of speech thus involves the violation of personal boundaries and psychological harm (Petlyuchenko et al., 2021). These two trends are sometimes offered as two dimensions. For instance,

¹ Perspective is an Application Programming Interface (API) that uses machine learning models to score the perceived impact a comment might have on a conversation. It reportedly processes over 500 million requests per day.

Waseem et al. (2017) argues abusive language can be categorized by taking into consideration the nature of the language used (implicit or explicit), and the target of the abuse (specific addressee or a generalized other). In the same vein, Kumar et al. (2018) propose a typology of verbal aggression by looking into both how it is expressed (overt or covert) and the nature of aggression (physical threat, sexual threat, identity threat, etc.) (see also Fortuna & Nunes, 2018).

The rationale in conceptualizing toxicity as offensive verbal behavior is that the more directly it is connected to a particular subject and an act of offence, the easier it becomes to normatively regulate it. Regulation and moderation of social media content have important roles to play in safeguarding pluralism in the online public sphere. Yet, the differences in size, reach, design, and business model of platforms are significantly involved in how content moderation works (Gillespie, 2020). This suggests the need for industry standards, but also a common understanding of the limits of admissible incivility, the delimitation of toxicity, and the regulatory enforcement agency.

The EU is particularly in favor of self-regulation by platforms, as they have the flexibility, agility, and innovation means to meet the evolving needs of online communities². The Code of Conduct published by the European Commission in 2016 represents an important step in this direction (Quintel & Ullrich, 2020). This initiative, adopted widely across platforms, requires that participating companies establish a set of rules and community standards explicitly forbidding online hate speech, submit such content for review, and remove it from platforms within 24 hours. The adoption of the Code also involved establishing a network of civil society organizations that monitors the implementation of these commitments (Reynders, 2022). Assessments of the Code reported impressive results in the number of processed user notifications, with a sharp increase in hateful content removal from the platforms from 2017 to 2020 (Reynders, 2022). Yet, serious concerns over lack of transparency and accountability remain, as little is known about how platforms process and remove content.

The process of content removal takes place in various ways, both internally by the platforms themselves, involving teams of humans and machine learning algorithms, and externally involving third-party companies. Notably, when content is flagged or reported by external experts, platforms have the final decision on removal. Platforms grant their users the option to report content, thereby leaving the detection of unwanted content to the community. A well-known case is the application of Reddit's 2015 anti-harassment policy, titled "Promote ideas, protect people"³, which caused many users to migrate from the

2 Platform companies are increasingly regarded as responsible parties such as the curators of the published content, rather than "mere conduits" or infrastructure providers (Mendis, 2021).

3 <https://www.redditinc.com/blog/promote-ideas-protect-people> (Accessed 30 Jan 2024).

platform. A serious limitation of user-based moderation is its subjectivity and openness to exploitation by user groups, which represents time and energy costs for platforms. Furthermore, during the processing of user reports, the content in question remains online, and this delay constitutes another limitation of user-based content moderation (Carrasco-Farré, 2022).

Another more technology-driven form of content moderation by platforms concerns the automated detection of toxic content, which involves machine learning algorithms. Indeed, machine learning and deep learning have been state-of-the-art in the last decade when it comes to hate speech detection (Jahan & Oussalah, 2023). AI-based systems proved highly effective at identifying certain content categories but were prone to errors with others (Ohol et al., 2023). AI-based systems came with large promises. Nevertheless, algorithmic moderation systems simultaneously suggest (a) further rises in the opacity of industry practices already lacking transparency; (b) exacerbating existing challenges regarding fairness within large-scale sociotechnical systems, and (c) depoliticizing inherently political decisions that might significantly influence public discourse (Gorwa et al., 2020). We revisit these limitations in the next section.

CONTROLLING ONLINE TOXICITY?

Despite the efforts mentioned above, controlling online hate and toxicity remains a difficult challenge. First, it is important to highlight the tension between uncivil language underpinning toxic debates and incivility integral to political expression. Even relatively nuanced forms of intervention based on a specific lexicon of “coarse language”, or predictive algorithmic content removal on any definition of toxicity, could stifle public debate. Disrespectful language may serve the minority or the discriminated groups who are otherwise not heard at all in public debate (Jamieson et al., 2017), and thus is integral to both the formation and makeover of public opinion. Incivility may also serve social purposes among like-minded people and be conducive to reasoned arguments (Chen et al., 2019; Rossini, 2019). In sum, inconsistent enforcement of cryptic standards across an industry consisting of competing corporations raises criticisms about suppressing dissident voices, which conflicts with the norm of freedom of speech (Quintais et al., 2023).

A second and associated drawback of lexically regulative approaches is that they operate at the micro-level of speech components, whereas cultural meanings and political implications often reside in the connections of a particular speech. For instance, words like shade, snowflake, or thirsty can be insulting across cultures, whereas slurs can be commonly used in non-toxic conversations (Sheth et al., 2022). Thirdly, regulating online interactions is largely at odds with

the makeup of the Internet. Conventional nation-state legislation and top-down enforcement will remain both spatially and temporally limited against the global reach and light-speed of the media. In comparison to broadcast or print media, the challenge is thus multiplied many times, requiring participation at various levels (Konikoff, 2021).

In sum, while content moderation efforts that operate at the micro-level of speech components help curb toxic commentary on social media to some extent, whether lexical matching or prediction-based (Gorwa et al., 2020), their impact on the online environments may be limited. There are positive steps that can be taken, which were briefly reviewed in the previous section. But, these efforts only scratch the surface of a more complex and multi-dimensional problem.

Rajadesingan, Resnick and Budak (2020, p. 559) argue that toxicity is not “an isolated phenomenon but a consequence of more structural factors” that have to do with each platform’s design and specific traits, content moderation policies, and community culture. In this regard, Oz et al. (2018, p. 3404) identify substantial differences between Facebook posts and tweets, with higher levels of aggression on Twitter (now known as X). They explain the difference by higher levels of de-individuation Twitter offers, as users communicate more often with strangers on this platform than on Facebook. Similarly, Recuero (2024) suggests that toxicity is fostered by the structural and economic particularities of platforms: “echo chambers” and “filter bubbles” are two of the famous metaphors that describe the users’ disconnect from the variety of available perspectives. The disconnect comes as a result of a customized information repertoire and “ideological homophily” fostered by platform algorithms, and its link to political polarization is well established (Boutyline & Willer, 2016).

TOXIC DEBATES: COLLECTIVE BUILDUP OF TROUBLED CONTEXTS

Following from the previous section, we hold that it is useful to distinguish broadly “toxic debates” from hate speech, abusive language, and toxicity. The argument is that toxic debates are not reducible to categories of speech by subject, but rather consist in an emergent feature of some polarized discussions. Feelings of hate and violence are sometimes collectively built, as suggested by the notions of “cascades” or “spirals” of toxicity (e.g., Kim et al., 2021). In this view, online toxicity is a socio-communicative issue with aspects that will escape all moderation – both by law and technology – and must be dealt with by platform users and communities.

The relationship between news topics and online toxicity is a case in point. Some issues are more controversial than others, and thanks also to media ranking algorithms, more divisive for societies (Milli et al., 2023; Recuero, 2024).

Research into online toxicity shows that a significant part of troubled comments is directed to the topic rather than individual users or groups, and that levels of toxicity vary significantly between topics (Salminen et al., 2020). Accordingly, topics with political connotations are more divisive for the online community, and topics such as the environment, health, race, and religion generate more hostile user comments. In turn, users who comment frequently on Facebook are shown to exhibit higher levels of political interest, possess more polarized viewpoints, and are more prone to employing toxic language in an elicitation task (Kim et al., 2021). For Salminen et al. (2020), this “topic-driven toxicity” suggests the potential impact that topic selection and the framing of news stories have on the shape and quality of social media discussions. Thus, as Salminen et al. (2020) argue, journalists today have additional burdens, since they should “be aware of the content topic’s inflammatory nature and possibly use that information to report in ways that mitigate negative responses” (p. 17).

Similar concerns also burden politicians, civil society organizations and platform users. We hold that the achievement of enduring results in curbing online toxicity relies on bottom-up understanding by, and the participation of, users. As in any democratic undertaking, sufficient emphasis should be placed on moral agency and online cultures. In this regard, it is no surprise that the notion of “netiquette”, the first informal code of online conduct (Kleinsteuber, 2004), appeared long before the soft laws and regulations that have entered the scene in the last decade.

However, we also need to recognize the role of “de-individuation” in the dynamics underlying toxic debates. Characteristics of online communication such as lack-of-face interaction, anonymity, and virtually instant access to unprecedented distances and audiences play a role in cascading toxicity. One aspect of this concerns the new speech context social media platforms provide for people to express themselves more freely than in other settings, a phenomenon dubbed the “online disinhibition effect” (Suler, 2004). Another very much interlinked aspect concerns the propagation or contagion of toxicity on media platforms. In this regard, Kim et al. (2021) identify amplification, mimicry, and normativity as three mechanisms that produce “spirals of toxicity” (p. 7). This spiraling effect of contagion is also documented in online gaming platforms (Shen et al., 2020).

This suggests asking the extent to which anyone can rely on individual users in the age of algorithmic concealment, celebritization, and the erosion of the contextual dimension of communication, where users find themselves “placed before random influences without knowing what they are, nor where they come from” (Cardoso, 2023, p. 47). How do users perceive and think about their regular experience with toxic encounters? What are their main worries and imaginaries of their future interactions online? We know too little about how platform users

consider toxicity, their views on what should be done, and the responsible agency. Scant research focuses on the perceived degrees of severity of the types of norm violations (Bormann, 2022), and the interaction with variables such as gender and political affiliation (Madhyastha, Founta & Specia, 2023).

Answers to any of the questions above contemplate as common responsibility the containment of a “global information environment crisis” (IPIE, 2024). We emphasized the political and cultural aspects of this responsibility, when with 29 assorted experts participating in workshops (see Table 1 of the the Introduction of this Special Issue and also Table 1 of this paper) and an essay writing project involving the 6 authors of this paper, we co-created 41 scenarios. We analyzed the scenarios to identify salient patterns and insights for thinking about the futures of networked communication, as presented in the next section.

FUTURE SCENARIOS ON TOXIC DEBATES

As the Introduction explains, “toxic debates and pluralistic values” comprised one of the five themes covered in the four Delphi+ workshops, which the EUMEPLAT team analyzed (see Table 1 below).

Table 1. The Delphi+ workshops, scenario-building exercises and theme specific codes for ‘Toxic Debates’ [txd]

Delphi+ workshops			
Locations	Codes and (frequency of) Scenario Cards	Participant code (Pn) in the pertinent location*	Theme and Location Specific Scenario Cards: SC[txd]n
Sofia 1	Si (7)	P2	SC[txd]1 – 7
Malmö	M (9)	P3, P4	SC[txd]8 – 16
Rome	R (10)		SC[txd]17 – 26
Sofia 2	Sii (7)	P1	SC[txd]27 – 33
Total	33	29	33
Future Scenario Essays			
	Number of Future Scenario Essays		Theme Specific Future Scenario Essays: FSE[txd]n
	8		FSE[txd] 1-8

Key: * There were 29 participants of the workshops: Si (6); M (6); R (7); Sii (10) (see Table 1 in the Introductory article of this Special Issue). These four participants—P1, P2, P3 and P4— were those cited in this article

We carried out content analyses (both quantitative and qualitative) informed by a narrative approach⁴ that pays attention to the pentad of actors, acts, scenes, agencies and purposes (Burke, 1969; Hänninen et al., 2022). Narrative may be regarded as a conventional mode through which people process and structure information (Bruner, 1991), as well as a human cultural effort to transform the feelings associated with certain events into a coherent sequence to learn from them (van den Hoven, 2017). In this view, narratives have an evaluative aspect, created through the connection of two casualties: a precedent event – complication – changes the circumstances of an actor, requiring her to respond, creating a succeeding causality. The succeeding action – repair – is central as it establishes the causal sequence that helps to construct the experience and drive lessons (van den Hoven, 2017). Futures studies seek to identify recurrent themes that tell us something about the underlying patterns that shape how people understand the future in a causal framework (Inayatullah, 2008). The narrative approach is useful in offering structure to what otherwise might be rather disconnected comments on the future.

The unit of analysis was the scenario, and our coding grid included the following nine fields: Title; Question(s) raised; Scene in the background; Main actor (of significant change), Main event (about Toxic Debates); Value (that grounds the aspired or unwelcome future); Prescription⁵; Role of the EU; Pessimism/Optimism. Except for the last field, all the others were coded by following an inductive approach. That is, rather than imposing top-down categories, we first coded particular actors, events, etc. Once the initial coding was finished, we grouped these figures into simple categories (e.g., human actors vs. non-human actors), and where necessary, into further, more diversified sub-categories.⁶

In the phases of categorization, we tried to remain attentive to the common patterns and causalities that weave the coded content together. The concept of myth (Inayatullah, 2008) was used to summarize these patterns and causal relations that connect the present to the futures envisaged in the scenarios. The Causal Layered Analysis for futures thinking (Inayatullah, 2008) stipulates myth

4 Drawing on Burke's dramatisic pentad (1969) and inspired by its relation to the study of futures (Hänninen et al., 2022), we initially attempted a narrative analysis, but encountered several difficulties. Some fields (Scene, Main Event, Agency) could not be coded systematically and had to be excluded from the analysis. This was because the scenarios differed considerably and were too sophisticated for this type of coding. Therefore, we opted to carry out quantitative and qualitative content analyses.

5 Again, inspired by Burke's (1969) dramatisic pentad, Main Event translates the Act into an action that is not necessarily connected to a particular actor, Value translates Purpose along the same lines, while Prescription registers the lesson – the coda, epilogue – that the narrative offers.

6 Given the very basic nature of the quantitative coding, and the limited number of texts, we decided against the calculation of an intercoder reliability coefficient. Instead, the author team checked the quality of the coding.

as “the deep unconscious story” (p. 12), akin to master narratives (Hyvärinen, 2020). We assume that while myths, like master narratives, have a taken-for-granted and archetypal character (Coward, 2022; Inayatullah, 2008), they can be disclosed, expressed and challenged (Hyvärinen, 2020). It is indeed a strength of futures studies to make explicit the visions of the future in a way that acknowledges not just the restrictive but also the productive power of such cultural stocks of stories (Hänninen et al., 2022). We thus use myths as cultural and communicative resources that people draw on when discussing possible futures, and as an interpretative tool to weave the content together, consisting of the causal connections among the common patterns and storylines. In the following sections, we report our quantitative and qualitative findings.

“EDUCATE PEOPLE, NOT MACHINES!”

ACTORS

We start the overview of the scenarios departing from the most relevant code in understanding the agency involved in constructing the futures of toxic debates and pluralism: actors. This code aimed to register the actor (actant) that brings significant change in each scenario. The code was split into three actor categories (Table 2), besides the null category—No actors identified—included those instances where a passive voice dominated the conversation, e.g., “Everybody will be anonymized. [...] Like the memes you lose track of everything” (SC(txid)16)⁷.

Table 2. Main Actor Categories

Actor	N
Digital and technological	19
Political and institutional	9
Media	5
No actors identified	8
TOTAL	41

The outstanding finding in this code concerns the predominance of non-human actors (19 of 33 scenarios with an actor mentioned), which are specifically digital or technological actants, such as “chatbots”, “artificial intelligence”, “algorithms”, “interface”, “platforms”, “journalistic machines”, “WeChat”, and “technology” at large. This predominance may be an outcome of the hype built around the

⁷ The number refers to the specific scenario card, see Table 1.

rise of generative AI at the time of the workshops. It simultaneously indicates the preoccupation of the participants with the enormous social impacts of recent developments in the computational sciences.

Following on from the dominance of non-human actors are 14 human actors, of whom 9 are political and institutional and 5 are media (see Table 2). Among the institutional and political agency, we can distinguish: “right-wing and populist parties”, “alternative and marginalized voices”, “colonizers”, “the acceleration”, “the public”, “Europe”, “media literacy programs”, and “some authority”. The term “colonizers” was used for denoting the human actors behind the algorithms regulating public opinion and human consciousness.

The five occurrences of the Media category are “Media”, “Niche media”, “Fake news” (twice) and “Public Service Media Organizations”. Note that fake news is a category that partially belongs to the political domain, due to being often used by illegitimate political interests. Without these two occurrences that pertain to pessimistic scenarios, niche media and public service media stand out as the sole actors that are set to bring some change from the media domain to the transformation of toxic communication.

VALUES

We report the values that pertain to communication and that the scenarios explicitly take up. These values typically ground the discussion over the imagined futures, more precisely, the actions and impacts brought about by the key actors, and they can be grouped into four categories (Table 3).

Table 3. Categories for Values

Values	N
Intellectual	13
Ethical	14
Sociopolitical	8
Technological	3
No values identified	3
TOTAL	41

In some contrast with the code Actors, values related to technology occupy a very small place in the scenarios. Instead, Intellectual (13 of 38 scenarios with a value mentioned) and Ethical (14 scenarios) take precedence in the futures of toxic communication and pluralism. To better understand these, we can exemplify Intellectual values as follows: “critical thinking”, “media critical thinking”, “media literacy”, “solid starting points”, “tolerance comes from knowledge”, and “substance of debate”. Notably, there were no negative values

among those that relate to the intellect, suggesting the participants' interest and esteem in the powers of reflection in tackling toxicity and a view of pluralism as an intellectual virtue.

Ethical values occupy a significant place in the scenarios and can be exemplified by “peaceful communication”, “respect”, “tolerance”, “pluralism”, “identity politics”, and “sensationalism”. The latter two of these are negative values in the sense that they are related inversely to pluralistic values and regarded as contributing to toxic debates. For instance, in one case (SC(txd)10), future generations, who live in the “vibe” of cancel culture and “social media as constant performative purity test”, fall prey to a “sort of compartmentalized identity politics”. It is then this negative vibe that brings about their failure in reconciling the two contradictory goals of “free speech” and “protect people from speech”.

In the third place are the values we designated as Sociopolitical (8 scenarios), a minority of which were negative. While “public good”, “transparency”, “universal citizen rights”, and “legitimate authority” are considered as positive values, “corporate interest” and “authority” exemplify the negative values.

Finally, the three occurrences of Technological values can be captured as “autonomy of technology”, “lack of face communication” (in online communication), and “digital mobility” (between bubbles, as a capacity that is achieved technologically). Notice that the initial pair are negative values – with autonomy of technology referring to the loss of human control over technological change. This suggests that when technology is linked to values grounding decisions or actions, it does so rather negatively.

PRESCRIPTIONS

This code aimed to register the policy proposals the scenarios may involve. It is typical of the pessimistic scenarios, in the sense that most of them devise an issue or problem – e.g. deep bubbles, the demise of the notion of truth – and then offer certain ways out of the predicament. A total of 25 of 41 scenarios involved such ideas towards positive change, or prescriptions. We initially coded these into two categories, which reflected the two fundamental aspects of social change—structure and agency (Best, 2014). The output of the coding process was rather unexpected, with all but one of the prescriptions being categorized as ‘structural change’ (24 scenarios). Building on the previously reported codes, we re-coded ‘structural change’ to distinguish it from the prescriptions that centrally involved technology. This way we achieved three categories for the code prescriptions (Table 4).

Even after the attempt to split the structural change code into two, there is still an overwhelming weight of structural change prescriptions (21 of 41 scenarios). This reflects the locus of deliberate change and social transformation as pointed out by the participants. Rather than individual or ethical action prescriptions—except for one

case—all scenarios involving such action-guiding proposals expected the change to originate in the structure, i.e., institutions and regulations, as these examples show:

“...Yes, encouraging pluralism. So, first to distinguish what are the hidden forms of dialogue that we can encourage and then to provide the tools for the people to be able to participate with them, because, the first one is how they can break this you and me contradiction model” (P1 at Sii).

“An obligatory continuous media education is implemented in schools of all types [...] The compulsory information and media education is a part of educational systems among Europe in all stages of education” (FSE(txid)5).

In the first of these two excerpts, the participant aligns themselves with a top-down agenda that provides tools for the public, encouraging novel formats of dialogue. The second excerpt also exemplifies the scenarios in which the “encouragement” is envisaged in a more structured educational reform. Such a position echoed in most of the scenarios, where education at large, and “encouraging activism, finding other way[s] to [...] participation” (SC(txid)39), or “democratization of culture and knowledge worldwide, and algorithm knowledge” (SC(txid)24), were offered as the locus of the solution(s) to online communicative predicaments.

Table 4. Categories for Prescriptions

Prescriptions of change	N
Structural	21
Technological	3
Agential and personal change	1
No prescriptions identified	16
TOTAL	41

To emphasize the weight of digital literacy and education in prescriptive statements, more examples can be offered. One of the scenarios elaborated several levels of intervention (FSE(txid)5): first, development of critical thinking for evaluating (online) content; second, encouraging empathy and respectful online interactions; third, encouraging responsible digital citizenship; and fourth, addressing online hostility. In another, we have critical perspectives in education: “...very close to this critical thinking. Progress through education, consensus through education and through developing critical thinking” (P2 at Si). Such calls for “progress through education” should not be regarded as un-reflexive prescriptions of simple modernization, as participants are well aware of the limits and failures of education as a policy to deal with social problems. Yet, they seem to be unable to come up with alternative proposals, probably due to the recognition of the necessity to approach such communicative problems in a bottom-up fashion.

To a lesser extent than the prescriptions on what may be called ‘critical thinking’ and ‘digital literacy’, there were others for more and extensive ‘regulation’. These were typically top-down measures to control and restrain the corporate power dominating in social media platforms and networked communication. Examples are “Regulation of commercial platforms” and:

“...interventions in business models, aligning with democratic principles [...] platforms cannot be operated with the same profit margins as before [...] Political support must come both from the nation-states and from the European level” (FSE(txid)2).

The need for regulation is recognized as an integral task for nation-states. Rather than imagining some new and innovative agency, for instance at the global level – except for “good algorithms” – the recorded prescriptions ascribed responsibility to current public authorities and governments. This seems to suggest that for the participants toxicity is a problem to be dealt with and a phenomenon that can be regulated today, rather than in an imagined future.

After examining the prescriptions for structural interventions, let us also briefly look at the outlier: the only scenario that included aspects of agential/personal change as a response to the bleak futures of online debates. This prescriptive statement also comprised algorithmic knowledge and digital literacy:

“P3 at M: [A] lot of people are gonna be like: I’m done having choices made for me, you will have to extricate yourself from a lot of systems” [...]

P4 at M: I also think that [this has] something to do with media literacy as well [...] so maybe the flip side is not just being offline or AFK [away from keyboard], but actually learning more about how things work, like how algorithms for how media works and so forth...”

It is worth noting that while the source of salvation is the same as with the majority of the prescriptions marked just above, in this excerpt the predicate is to “learn” – rather than “encourage” – and it signals the powers or agency of the user in a bottom-up fashion. While it plays the agential tune in reverse, in regard to the content, the outlier also falls within the broad domain of digital literacy, with an emphasis on acquisition and self-instruction on how algorithms work.

In brief, two major messages can be drawn from the prescriptive statements involved in the scenarios analyzed: educate and regulate. In this regard, perhaps the most salient direction that can be drawn from the experts involved in the scenarios is summarized in a slogan that popped up in one of the sessions: “Educate people, not machines” (SC(txid)20).

As an epilogue to this section, let us briefly mention the role of Europe in the scenarios. Europe was mentioned only in 11 of 41 scenarios. Although it was hardly one of the central actors, it was endowed with a consistent character, namely with the role to “safeguard democracy”, “defend the institutions” (FSE(txd)1), and “among the institutions most likely to foster, and cultures most prepared to sustain, such an open public debate” (FSE(txd)6). The EU was thus ascribed a central role in the public education and digital literacy efforts mentioned above: “Under the coordination of European institutions, specific modules designed to combat toxicity could be established in schools” (FSE(txd)8). Besides this, there were also few mentions of a “stronger European identity”, and, more precisely, the recommendation “the EU should empower its tech and media industry to take the lead, even to import know-how from abroad, since most European AI companies are still at an early stage” (FSE(txd)7). Generally speaking, the EU was not a defining actor in the scenarios, but there were calls for it to become one, if toxicity and fragmentation of society were to be tackled.

ENVISAGING THE FUTURES OF TOXIC DEBATES

In more or less organized ways, societies increasingly project themselves into the future, set goals, and strive to contain the externalities of their preceding projections. Future, in this sense, becomes a resource to orient human action and policy preferences (Üzelgün & Pereira, 2020). After the study of future scenarios, we now interpret the coded categories to extract the salient causal relationships and myths from the 41 scenarios. This section discusses three myths and two causal relationships that characterize the scenarios, informed by the quantitative content analysis, and further supported by a qualitative content analysis.

■ TECHNOLOGICAL DISRUPTION

The first myth can be called technological, or more specifically, AI and algorithmic disruption. It underlies the imaginaries of a brave new world where the integration of digital technologies into all aspects of human communication brings numerous challenges that even the public cannot fully comprehend. This myth is grounded in the overwhelming predominance of the AI and digital actants among the actors that bring the change, as well as that almost no agency is ascribed to the user or the public in the prescriptive statements. In other words, the most central preoccupation of the participants was that digital and generative technologies bring a sweeping change that will disrupt manifold aspects of human communication. Rapidly evolving digital technologies are thus envisaged as the villain and the main cause of future predicaments. Yet, to address

how these technologies impact and interact with toxic debates, it is imperative to understand how they broadly tap into the mechanisms of “virality” and platform logics (Recuero, 2024).

■ SOCIETAL FRAGMENTATION

The second myth can also be called by its sociological metaphor—anomie. As the corporate deployment of algorithms, AI and other disruptive technologies amplify existing cleavages, nothing short of the breakdown of common grounds and communicative frameworks is regarded as the peril ahead. Societal fragmentation thus consists in the communicative predicaments online, summarized in the notion of toxicity, but exacerbated by technology as projected into the future. This central myth then represents where the scenario builders envisage themselves with regard to toxic debates: a broken society that could not anticipate the social and political impacts of the disruptive technologies mentioned above. Several cascading factors and issues can be aligned in this causal link: lack of facework, filter bubbles, fake news, polarization, blurring boundaries of the real and virtual, and the neglect of truth. In short, regarding platformized interactions, designed and maintained by non-human values and interests, the central worry is the loss of the foundational elements of human communication, remaining locked in conflicts and contradictions.

■ ENLIGHTENMENT 2.0

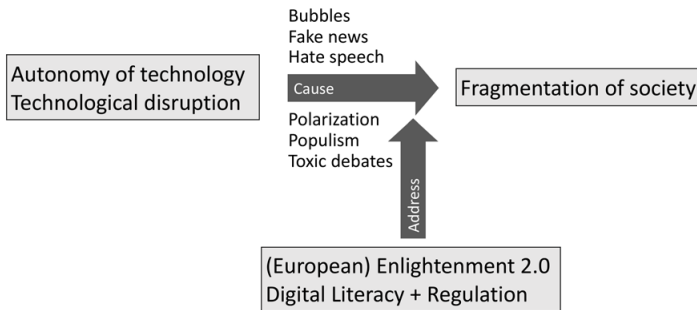
The third myth is associated with Europe and consists in a decidedly digital enlightenment – hence the 2.0 designation – in which authorities are envisaged to encourage digital literacy, public knowledge on algorithms, critical thinking to evaluate online information, and support the epistemic quality or substance that grounds public debate. Notably, enlightenment 2.0 is not just about enhanced critical thinking on the part of users, but also about regulating the platforms and the corporate interests behind algorithmic distortion. In this sense, a core concern is public – or human – accessibility, and corporate accountability, of the choices taken over digital platforms. The regulations mentioned also concern upholding and innovating in public service media, opening alternative paths to media institutionalization, and innovation in the design of online debate and interactions. Enlightenment 2.0 thus incorporates both bottom-up and top-down measures to address as yet little known impacts of platformization on human communication and society. Although it was not as salient, digital enlightenment represents the collective efforts envisaged to deal with the communicative predicaments registered by the previous two myths, and has an important role in the construction of futures.

■ CAUSAL LINKS

To address the relations among the three myths that summarize the futures of toxicity, two causal relationships may be discussed (See Figure 1). The first causal link lies between the first two myths, depicting the challenges that digital technologies precipitate for the complex communicative predicaments captured by the notion of toxicity. This means, issues such as filter bubbles and polarization are projected to exacerbate with further development of platform technologies. The impacted end of the causal link is human society at large, and an associated worry is that the public may not be ready to handle, nor comprehend, the challenges human nature and institutions are faced with.

It is important to underline that, contrary to what Figure 1 may suggest, technology is not the only cause that brings about the second myth—fragmentation of society. Technology should be seen as exacerbating the already existing societal problems. In this sense, the loci of the relationships among the three myths are the six problems that connect all three imaginaries: bubbles, fake news, hate speech, polarization, populism, and toxic debates.

Figure 1. A basic view of relationships among the three myths



If the link between the first and the second myths was causal, that between the second and the third myths could be designated as negative causation. That is, the third myth is envisaged to avert the impact of digital technologies on society, by protecting communicative and social relations. In other words, to address the ongoing fragmentation of society due to the platform designs, the recommendation is to launch a global public campaign to enhance digital literacy and regulate social media platforms, with the ultimate objective of boosting democratic accountability. In this regard, calls for regulation, associated with the institutional level, may be seen to indicate a certain concern, or fear of the AI-powered algorithmic distortion as a “symptom” of deregulation and neoliberalism.

TRANSFORMING PLATFORMIZED INTERACTIONS

Animated by platform monetization and recommendation algorithms, toxicity endangers not only pluralism and quality of societal debates (Anderson et al., 2018; Milli et al., 2023), but also the future of public discourse at large. The pessimistic tenor of the scenarios examined in this paper, and specifically the dim view of the role of technology therein, can be understood within the framework of a loss in the media gatekeeping processes (Cardoso, 2023). As the static gatekeeping practices are transformed into a dynamic practice of negotiation between users and algorithms (Cardoso, 2023; Konikoff, 2021), the aspects that becomes increasingly invisible and unintelligible are the rules of the negotiation. The lack of transparency and social understanding of the network gatekeeping processes may account for absences in the scenarios of a view favoring the injection of democratic values into these dynamic processes, as well as that of fostering participatory self-regulation by users (de Gregorio, 2020). So, concerning the futures of toxic debates, the complex challenge ahead can be simplified twofold. First, platform transparency, which rather than optimistically expected from platform businesses, should be imposed as a public good. Second, as a much more complex challenge, empowerment of online users, communities and initiatives to actively participate in the vast potential opened by digital technologies. After all, the future of the networked debates will probably depend on the extent to which we understand who keeps the gate and how.

ACKNOWLEDGEMENT

This article is part of the EUMEPLAT project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004488. Any dissemination of results and any communication must indicate that it reflects only the author's view and that the Agency is not responsible for any use that may be made of the information it contains.

REFERENCES

- Adams, C.J., Sorensen, J., Elliott, J., Dixon, L., McDonald, M., Nithum, & Cukierski, W. (2017). Toxic comment classification challenge. <https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge>
- Anderson, A. A., Yeo, S. K., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2018). Toxic talk: How online incivility can undermine perceptions of media. *International Journal of Public Opinion Research*, 30(1), 156-168. <https://doi.org/10.1093/ijpor/edw022>

- Best, S. (2014). Agency and structure in Zygmunt Bauman's modernity and the holocaust. *Irish Journal of Sociology*, 22(1), 67-87. <https://doi.org/10.7227/IJS.22.1>
- Bormann, M. (2022). Perceptions and evaluations of incivility in public online discussions: Insights from focus groups with different online actors. *Frontiers Political Science*, 4:812145. <https://doi.org/10.3389/fpos.2022.812145>.
- Boutyline, A., & Willer, R. (2016). The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology*, 38, 551-569. <https://doi.org/10.1111/pops.12337>
- Bruner, J. (1991). The narrative construction of reality. *Critical inquiry*, 18(1), 1-21.
- Burke, K. (1969). *A Grammar of Motives*. University of California Press.
- Cardoso, G. (2023). *Networked communication: People are the message*. Mundos Sociais.
- Carrasco-Farré, C. (2022). The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications*, 9(1), 1-18. <https://doi.org/10.1057/s41599-022-01174-9>
- Chen, G. M., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media + Society*, 5(3), 1-5. <https://doi.org/10.1177/2056305119862641>.
- Cowart, A. (2022). Living between myth and metaphor: Level 4 of Causal Layered Analysis theorised. *Journal of Futures Studies*, 27(2), 18-27. [https://doi.org/10.6531/JFS.202212_27\(2\).0003](https://doi.org/10.6531/JFS.202212_27(2).0003)
- de Gregorio, G. (2020). Democratising online content moderation: A constitutional framework. *Computer Law & Security Review*, 36, 105374. <https://doi.org/10.1016/j.clsr.2019.105374>
- Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM*, 57(6), 74-81.
- Ferrara, E., & Yang, Z. (2015). Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science*, 1(26), 1-15. <https://doi.org/10.7717/peerj-cs.26>
- Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4), 1-30. <https://doi.org/10.1145/3232676>.
- Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2), 1-5. <https://doi.org/10.1177/2053951720943234>.
- Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 1-15. <https://doi.org/10.1177/2053951719897945>
- Guberman, J., Schmitz, C., & Hemphill, L. (2016). Quantifying toxicity and verbal violence on Twitter. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW '16 Companion)*, 277-280. New York: Association for Computer Machinery.
- Hameleers, M., Bos, L., & de Vreese, C. H. (2017). "They did it": The effects of emotionalized blame attribution in populist communication. *Communication Research*, 44(6), 870-900.
- Hänninen, V., & Sools, A. (2022). Cultural story models in making sense of a desired post-corona world. *Futures*, 141, 102989. <https://doi.org/10.1016/j.futures.2022.102989>
- Hyvärinen, M. (2020). Toward a theory of counter-narratives: Narrative contestation, cultural canonicity, and tellability. In K. Lueg & M.W. Lundholt (Eds.), *Routledge handbook of counter-narratives* (pp. 17-29). Routledge.
- IPIE, International Panel on the Information Environment (2024). <https://www.ipie.info/> (Accessed January 30, 2024).

- Inayatullah, S. (2008). Six pillars: Futures thinking for transforming. *Foresight*, 10(1), 4-21.
- Jahan, M. S. & Oussalah, M. (2023). A systematic review of Hate Speech automatic detection using Natural Language Processing. *Neurocomputing*, 546. 126232. <https://doi.org/10.1016/j.neucom.2023.126232>.
- Jamieson, K. H., Volinsky, A., Weitz, I., & Kenski, K. (2017). The political uses and abuses of civility and incivility. In K. Kenski & K. H. Jamieson (Eds.), *The Oxford handbook of political communication* (pp. 205-218). Oxford University Press.
- Jigsaw (2024). <https://current.withgoogle.com/the-current/toxicity/> (Accessed February 7, 2024).
- Kim, J. W., Guess, A., Nyhan, B., & Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, 71(6), 922-946. <https://doi.org/10.1093/joc/jqab034>
- Kleinstueber, H. J. (2004). *The Internet between regulation and governance. Self-regulation, co-regulation, state regulation*. <https://www.osce.org/files/f/documents/2/a/13844.pdf> (Accessed on 14 February 2024).
- Konikoff, D. (2021). Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies. *Policy & Internet*, 13(4), 502-521. <https://doi.org/10.1016/j.clsr.2019.105374>
- Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of the workshop on trolling, aggression and cyberbullying, 1-11*.
- Madhyastha, P., Founta, A., & Specia, L. (2023). A study towards contextual understanding of toxicity in online conversations. *Natural Language Engineering*, 29(6), 1538-1560. <https://doi.org/10.1017/S1351324923000414>
- Mendis, S. (2021). Democratic discourse in the digital public sphere: Re-imagining copyright enforcement on online social media platforms. In H. Werthner, E. Prem, E. A. Lee, & C. Ghezzi (Eds.), *Perspectives on Digital Humanism* (pp. 41-46). Springer.
- Milli, S., Carroll, M., Pandey, S., Wang, Y., & Dragan, A. D. (2023). Twitter's algorithm: Amplifying anger, animosity, and affective polarization. *arXiv preprint arXiv:2305.16941*.
- Neff, G. (2024). The new digital dark age. *Wired*. <https://www.wired.com/story/the-new-digital-dark-age/> (Accessed February 23, 2024)
- Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web*, 145-153. <https://doi.org/10.1145/2872427.2883062>.
- Ohol, V. B., Patil, S., Gamne, I., Patil, S., & Bandawane, S. (2023). Social shout: Hate speech detection using machine learning algorithm. *International Research Journal of Modernization in Engineering Technology and Science*, 5, 584-586.
- Oz, M., Pei, Z., & Chen, G. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New Media & Society*, 20, 3400-3419.
- Pascual-Ferrá, P., Alperstein, N., Barnett, D. J., & Rimal, R. N. (2021). Toxicity and verbal aggression on social media: Polarized discourse on wearing face masks during the COVID-19 pandemic. *Big Data & Society*, 8(1), 1-17. <https://doi.org/10.1177/20539517211023533>
- Patel, A., Cook, Ch. L., & Wohn, D. Y. (2021). User opinions on effective strategies against social media toxicity. *Proceedings of the 54th Hawaii International Conference on System Sciences*. <http://hdl.handle.net/10125/70980>

- Petyuchenko, N., Petranova, D., Stashko, H., & Panasenکو, N. (2021). Toxicity phenomenon in German and Slovak media: Contrastive perspective. *Lege Artis. Language Yesterday, Today, Tomorrow*, 2, 105-164.
- Quintais, J. P., De Gregorio, G., & Magalhães, J. C. (2023). How platforms govern users' copyright-protected content: Exploring the power of private ordering and its implications. *Computer Law & Security Review*, 48, 105792.
- Quintel, T., & Ullrich, C. (2020). Self-regulation of fundamental rights? The EU Code of Conduct on Hate Speech, related initiatives and beyond. In B. Petkova & T. Ojanen (Eds.), *Fundamental Rights Protection Online* (pp. 197-229). Edward Elgar.
- Rajadesingan, A., Resnick, P., & Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1), 557-568. <https://doi.org/10.1609/icwsm.v14i1.7323>.
- Recuero, R. (2024). The platformization of violence: Toward a concept of discursive toxicity on social media. *Social Media + Society*, 10(1). <https://doi.org/10.1177/20563051231224264>
- Reynders, D. (2022). *7th evaluation of the Code of Conduct on countering illegal hate speech online*. European Commission. https://commission.europa.eu/document/download/5dcc2a40-785d-43f0-b806-f065386395de_en?filename=Factsheet%20-%207th%20monitoring%20round%20of%20the%20Code%20of%20Conduct.pdf
- Rossini, P. (2019). Toxic for whom? Examining the targets of uncivil and Intolerant discourse in online political talk. In P. Moy & D. Mathlson (Eds.), *Voices: Exploring the shifting contours of communication* (pp. 221-242). Peter Lang.
- Salminen, J., Sengün, S., Corporan, J., Jung, S., & Jansen, B. J. (2020). Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLoS ONE*, 15(2): e0228723. <https://doi.org/10.1371/journal.pone.0228723>
- Shen, C., Sun, Q., Kim, T., Wolff, G., Ratan, R., & Williams, D. (2020). Viral vitriol: Predictors and contagion of online toxicity in World of Tanks. *Computers in Human Behavior*, 108, 106343. <https://doi.org/10.1016/j.chb.2020.106343>.
- Sheth, A., Shalin, V. L., & Kursuncu, U. (2022). Defining and detecting toxicity on social media: Context and knowledge are key. *Neurocomputing*, 490, 312-318. <https://doi.org/10.1016/j.neucom.2021.11.095>.
- Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and information diffusion in social media sentiment of microblogs and sharing behavior. *Journal of Management Information Systems*, 29(4), 217-248. <https://doi.org/10.2753/MIS0742-1222290408>.
- Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, 7(3), 321-326.
- Üzelgün, M. A., & Pereira, J. R. (2020). Beyond the co-production of technology and society: The discursive treatment of technology with regard to near-term and long-term environmental goals. *Technology in Society*, 61, 101244. <https://doi.org/10.1016/j.techsoc.2020.101244>
- van den Hoven, P. (2017). Narratives and pragmatic arguments: Ivens' The 400 million. In P. Olmos (Ed.), *Narration as argument* (pp. 103-121). Springer Cham.
- Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, 78-84. Association for Computational Linguistics.
- Williams, J. (2018). *Stand out of our light: Freedom and resistance in the attention economy*. Cambridge University Press.