CELE
Centro de Estudios en Libertad de Expresión y Acceso a la Información

Glossary of Trust & Safety Terms

CELE's Submission

March 15, 2023

The present consultation is calling for feedback to the Glossary of Trust and Safety Terms, in the context of a "discipline" that is growing. The terms defined in the document, thus, are useful to assess the state of said discipline, to inquire into what practitioners and researchers working within it believe is the scope of the approach, what is included, and what is left outside the reach of this rather new field of inquiry and practice. We thus approach the task ahead with an open mind but with questions that we consider useful to share, in order to encourage the self-reflection that the calls for comments seeks to promote.

Our first, general comment would pose that we are unsure what Trust and Safety means. Taking the document as a point of departure, it is defined as a *discipline*, a way in which we researchers generally describe a field of inquiry, where researchers share concerns, address similar problems, pose questions, approach the object of study from different methodological approaches, and so on. From that standpoint, the definition the document offers of Trust and Safety is useful. It is defined in page 8 as:

> "The field and practices employed by digital services to manage content and conduct-related risks to users and others, mitigate online or other forms of technology-facilitated abuse, advocate for user rights, and protect brand safety. In practice, Trust & Safety work is typically composed of a variety of cross-disciplinary elements including defining policies, content moderation, rules enforcement and appeals, incident investigations,

> law enforcement responses, community management, and product support. Since about 2005, it has developed into a distinct profession in its own right, with several professional organizations (such as DTSP and the Trust & Safety Professional Association) focusing on Trust & Safety functions emerging since 2020".

From this standpoint, Trust and Safety does not seem to be a discipline in the academic sense of the word: it is, rather, a way of grouping certain practices within private corporations who work as intermediaries in the information flow that happens through the Internet. We are all for it, but we would like to make the point: defining it as a discipline may project the wrong idea that the field is crossed by broad consensus on a set of shared concerns, the problems that must be addressed, and so on. We are not sure that is the case. In particular, we would like to highlight certain stress points that can be found in the definition, where disagreement is pervasive.

For instance, there is no agreement as to what counts as a *risk* for Internet users. While some may judge that being exposed to hateful speech is a harm that poses a risk (for the well-being of the user exposed to such content), others may deem that lack of exposure to said content as risky (because, e.g., speech that is not seen cannot be refuted, because a spiral of silence phenomenon may go unnoticed if a form of speech is repressed, and so on). Similarly, the idea that Trust and Safety advances *users rights* assumes that there is agreement as to what those rights are, but that is not the case. While some may feel they have a right to a platform that offers them a safe space for expressing their opinions and reading others', others may honestly believe that they have a right to offend other people (even through shocking, outrageous, and disturbing opinions, to follow international human rights law language). Both conceptions of the rights at stake in Trust and Safety are incompatible with each other. Which conception does the field embraces? *Law enforcement responses* also should be normatively assessed depending on the law-enforcement agency making the request (judges, administrative agencies, and so on) and the rule-of-law context in which law-enforcement agents operate (full-fledged democracies, weak democracies, hybrid regimes, authoritarian states, and so on). How are these differences

considered within the field? Finally, what does *brand safety* actually means? Is it that brands do not want to see themselves linked to problematic or controversial speech, to speech that is hateful or discriminatory, all of the above? What prerogatives do brands have to affect the way social media companies moderate content? What should they have?

This pervasive disagreement should be at the core of the reflection pushed forward by practitioners and researchers in a field that seems to be emerging. From these, even the fundamental shared concerns and questions that builds an autonomous discipline could be carved out. In that spirit, we highlight certain definitions that we find particularly problematic.

In page 4 community guidelines, community moderation, and content moderation are defined in the following way.

> **Community Guidelines**. The set of conditions and limitations governing use of a digital service that a user must agree to as a condition of use. These are generally written in plain and concrete language (compared to legal language used in terms of service). Also called "acceptable use policy," or content policies).
>
> **Community Moderation**. A method of content moderation whereby the users of a site or service (as opposed to site administrators or corporate employees or contractors) play a substantial role in reviewing and taking moderation actions on user-generated content.Community moderation may be a method of enforcing general sitewide community guidelines, or more specific rules or guidelines particular to a subpart of a service that the users have written independently. Community moderation has its origins in early-internet message board culture and is one of the oldest forms of online content moderation.
>
> **Content Moderation**. The act of reviewing user-generated content to detect, identify or address reports of content or conduct that may violate applicable laws or a digital service's content policies or terms of service. Content moderation systems often rely on some combination of humans and machines to re-

> view content or other online activity with automation executing simpler tasks at scale and humans focusing on issues requiring attention to nuance and context. The remedies resulting from violation of a service's policy can include disabling access to content, temporary or permanent account suspension, and demotion of distribution in search or recommendation engines, and other safety interventions such as those identified in Section III below.

We argue that the last two distinguish between two different forms of moderation, one in which the community takes part and one in which that participation is not necessarily in place (the former, presumably, encompasses the latter, it is just a specific form of content moderation). We welcome the distinction, but suggest that a similar nuance should be introduced to distinguish between the community guidelines that are defined by the community itself, from those sets of "conditions and limitations" that are established by company officials. We are aware that Internet companies usually speak of community guidelines to refer to company-set rules and principles, but this—in our opinion—misuses the rich texture of the concept of "community" within the Internet. In that sense, community guidelines should be better distinguished from terms of service.

On page 5, the Glossary defines "explicit content" in the following way:

> Online content describing or depicting things of an intimate nature. Depending on cultural context, this may include nudity, parts of the body not generally exposed in public, sexually explicit material, or depictions of sex acts. Sometimes used interchangeably with "adult," "intimate" or "NSFW" ("Not Safe for Work"), and may also include offensive, graphic, or violent content, or association with content or commerce involving gambling, sex, cosmetic procedures, recreational drug use.

We argue that this definition would not solve Facebook's nipple problem (e.g., how are we to distinguish between sexually explicit material and a campaign to raise awareness on breast cancer). This is a limitation that could be worked out in the definition.

On page 6, the right to be forgotten is defined in the following way:

> The right to be forgotten refers to the right of individuals to request that online services remove certain content related to them. For example, this may include the request to erase an individual's personal data and may apply where a search engine returns information that is inaccurate or irrelevant, and the publication of such information is not in the greater public interest. The concept is rooted in the concern that a single event, memorialized online, may impose unduly punitive consequences for a person's reputation, indefinitely, if there is no mechanism for reconsideration and removal. The right was first recognized in a 2014 ruling by the European Court of Justice, and was later codified as a "right to erasure" with the passing of the General Data Protection Regulation (GDPR) in 2018. A right to erasure has since been recognized in other jurisdictions, including Argentina, Russia, and the Philippines.

We note that it is a factual mistake to affirm that the right to be forgotten has been recognized in Argentina. It was done through an Appellate decision that was later overturned by the Supreme Court.

On page 9, the document defines brigading in the following way:

> Coordinated mass online activity to affect a piece of content, or an account, or an entire community or message board, for example by upvoting or downvoting a post to affect its distribution, mass-reporting an account (usually falsely) for abuse in an attempt to cause the service provider to suspend it, or inundating a business with good or bad reviews.

We consider it would be useful to define with more precision what "coordination" means in the context of the definition. Does it include spontaneous coordination, as in many people start to do the same thing that other people are doing at the same time or some degree of intent, planning, or organization is to be required to meet the definition? The point is—we believe—important because, generally, platforms should show much more tolerance to the former than to the latter.

On page 14, the document defines troll as a "user who intentionally provokes hostility or confusion online". We believe the definition of troll misses the mark of common usage (a person who is somewhat obnoxious or makes valid points in ways that are irritating is usually called a troll, but has no desire of provoking hostility or create confusion). At the same time, the concept of hostility should be in itself defined, in order to distinguish it from criticism.

Finally, on page 11 and 14 disinformation and misinformation are defined in the following way.

> **Disinformation**. False information that is spread intentionally and maliciously to create confusion, encourage distrust, and potentially undermine political and social institutions.
>
> **Misinformation**. False information that is spread unintentionally and usually not maliciously, which may nonetheless mislead or increase likelihood of harm to persons. (Compare with "disinformation.")

We consider that these definitions pose some problems. First, determining the "falsehood" of a statement appears to be a very hard task that, if carried out improperly, can have drastically adverse consequences on the openness of public debate. The second point is that, if disinformation is distinguishable from misinformation in that the former is intentional while the latter is not, then misinformation should not be placed under the "abuses" section of the document. Otherwise, a great deal of pieces might be subject to removal or flagging only because the content of the speech they convey is deemed "false" by a decision-maker, which could threaten the robustness of the public conversation.