ARPI

Australian Risk Policy Institute

Discussion Paper: Need for a Single, Global, Enforceable Standard on Social Media

---

In a world of accelerating innovation, increasing threats matched by reducing protection for the most vulnerable, social media has quickly moved from being an enthralling social opportunity and building social capital to devastating the lives, reputation and safety of innocent people. Disinformation and wrongful influence prevail. Innovation and changes or withdrawals of services without thought of consumer protection always starts as a major risk to society.

ARPI, the Australian Risk Policy Institute, an independent, non-profit organisation formed to promote the new concept of Risk Policy 4.0 to enhance leadership, decision-making and public policy, is today publishing a global frame for use of social media and protection of vulnerable people, in the public interest.

Four basic tenets should apply to a global Standard:

1. **Legislative coverage** for the strength, compliance and enforcement must occur. While presenting challenges in implementing overall global legislative coverage, we note that G20 is presently examining this point so a solution is understood to be needed, hence practicable;
2. A Standard must be a **scalable** with the size (and growth) of an organisation;
3. **Contracting out** must be prohibited – to prevent the creation of response-based processes that meet minimal budgetary outlays; and
4. As sole reliance on Artificial Intelligence (AI) for **Quality Assurance** has failed and will always fail in this environment due to inherent conceptual limitations e.g. abstract reasoning and situational awareness, it is necessary to prescribe the use of human involvement (through Intelligence Augmentation IA) to protect against vulnerabilities and avoid preventable, catastrophic failures.

Suggested Terms of Service (TOS) are as follows:

1. Users and Providers of the Service

Users and providers are bound by these terms of service.

2. Content

A user retains ownership of content they create and share to the service (including images, post text, comments, voice and video) notwithstanding sharing. However, the act of sharing gives the provider a licence to serve (re-publish) that content to other users on the service as agreed by the user when posting.

A provider may not strip copyright or creation metadata from an item (e.g, EXIF or IPTC on photo images) provided to the provider to be hosted on the service or converted for hosting on the service. All material shared on the service by the provider shall contain original

copyright or creation metadata with limited changes to reflect the creation of a new version, if any.
Where a user's data is hosted on the provider's site, the provider shall give reasonable support for the user to recover the user's property.

3. Truth
A user or provider shall honestly identify themselves and their home jurisdiction to each other, and a provider shall keep that information confidential.
A provider may agree to provide a user with an obviously fictive name for dealing with others on the Service. A user must not use a fictive name to mislead another user about the first user's real identity.
At all times, it is the responsibility of a person with a password in their possession to ensure safety of the password.

4. Equality, Respect and Free Speech
On the service, all are considered equal and deserving of respect. The provider shall act in good faith at all time demonstrating that the provider values respect to others and free speech.

5. Freedom from Criminal Activity
Conduct that would be a criminal offence in a user's home jurisdiction or any offence adjudged by most jurisdictions to be a crime against humanity (such as condoning or supporting dangerous drug importation, slavery or terrorism) will be muted by the provider.

6. Freedom from Scams, Trolling, Inciting Hate, Depicting Violence
Posts on the Service that incite hatred on the basis of race or ethnic origin, religion, disability, age, nationality, veteran status, sexual orientation, gender, gender identity or any other characteristic that is associated with systemic discrimination or marginalization will be muted by the provider.

7. Freedom from Political or Campaign Calls To Action
The posting of political or campaign material (such as a call to vote for a particular party or candidate, calls to action or recruitment to groups dealing with terrorist groups, climate or animals etc.) (but not including incidental material, problem solving screen shots or celebration of official holidays) should only be provided to people who have given prior agreement to receiving that material. Material in this group shall be muted by the provider if they are posted in a general stream.

8. Freedom from Commercial Activity
Advertising material for sale (such as inviting people to visit a site to view or purchase goods or services, selling artwork or books, or inviting people to join a service) (but not including incidental material, references to a personal website) should only be provided to people who have given prior agreement to receiving that material (e.g. members of a service funded by in-service advertisements or members who join a sub-community willing to consume ads). Material in this group shall be muted by the provider if they are posted in a general stream or if it appears to be related to drugs, hacking, spam marketing, adult dating, prostitution, or pornography.

9. Freedom from Threats
Harassing, threatening, sexualising (another person/themselves) or exploitative (e.g, the distribution of another person's intellectual property without consent) material shall be muted by the provider.

10. Reporting, Action and Suspension
The provider shall put in place such mechanisms as permits a user to effectively report any violation of these TOS.
A provider shall act quickly and may err on the side of safety of a user or the community when making a decision to mute provided that the provider reconsider the decision as soon as circumstances permit.
Where material is muted under these TOS, the provider shall consider whether to suspend the user and/or provide the muted material to police services together with details of the user responsible for posting the material.
While a user account is suspended, the user may not create a new account. The user may, within 7 days, appeal the suspension by writing to the provider setting out the reasons suspension is inappropriate and the provider shall giver natural justice to the user when considering the suspension.

11. Scalable Performance Standards
The provider shall put in place such mechanisms from time to time having regard to the size of the service as is appropriate to mute material at the earliest possible stage having regard to the danger imposed to an individual or the community at large. The services must not be contracted to an agency out of the provider, but may include appropriately supervised moderation by experienced users.

Comments, suggestions and feedback are invited to: administration@arpi.org.au

ARPI's Strategic Risk Policy Model is downloadable free at www.arpi.org.au

24 March 2019