The Diligent Platform and "Lawful but Awful" Expression

Emily Laidlaw

An online safety bill is expected to be introduced by the Government of Canada in the coming months. Regardless of what is proposed, this is just the beginning. The coming decade will be a period of tremendous legal friction as online safety laws are fine-tuned by courts, regulators, and legislators. Even if no law is passed, we will still have the same question of what precisely counts as a responsible platform to a legal standard. Several legal conundrums keep me up at night, and one more than all the others:

Can and should a risk management model target "lawful but awful" expression?

All indications from Heritage Canada are that the legislation will be modelled on what can be described as risk management, duty of care or due diligence models. Under these legal frameworks, the focus is on the systemic risks of harm of platforms, such as the recommender or other algorithmic systems, content moderation, advertising and data practices. These platforms' services are treated as issues of product safety. The obligation of the company, therefore, is to assess the risks of harm of their services at the front end, continually monitor the risks, and action any findings. Platforms are held accountable through a mix of mandatory transparency reporting and regulatory oversight. This approach is observable in the United Kingdom's *Online Safety Bill* (OSB), the European Union's (EU) *Digital Services Act* (DSA), and currently explored in Canada as a duty to act responsibly. All of these models, in essence, require that platforms implement systems of reasonable decision-making.

I have advocated for this model, but the analogy of product safety can only take us so far. Building safer cars is not the same as building safer spaces for discourse, because of the intersection with fundamental rights. The friction is most obvious when examining systems and content at the edges of legality.

When I discuss online safety legislation, most people mention mis- and dis-information, bullying, mob attacks, hateful content and similar. All this content is generally lawful but harmful. Can and should online safety legislation address this type of harm? This was an enormous point of controversy in the UK concerning the OSB and resulted in most such provisions removed from the bill. Still, there is an important conversation to be had about what platforms can and should be responsible within a constitutional framework.

For example, the 2022 coroner's inquiry in the UK into the suicide of a teen girl, Molly Russell, revealed that the recommender systems of both Pinterest and Instagram were pushing content on self-harm, suicide, and depression. A risk management model that only focuses on illegal expression would not demand due diligence by platforms in this sphere. However, a duty to protect the special interests of children and assess the risks of their recommender system would capture what caused harm to Molly Russell. Similarly, most mis- and dis-information is perfectly legal even if it might shape decisions linked with illegality. The EU's DSA creatively circumvents

the question of lawful but awful and disinformation by imposing due diligence obligations on very large online platforms (VLOPs) to manage their systemic risks without getting into the weeds of individual pieces of content or their legality.

Two legal challenges are key. First, the starting position for targeting any such content – even at a systemic level - is that a legal system of free expression protects unpopular, distasteful, and disturbing expression – and it should, because the foundations of democracy, discovery, and truth inherent in expression are central to society. However, such a right must be reinforced and balanced with other constitutional rights, like the rights to privacy and equality, the Crown's fiduciary duties to Indigenous people, and the unique infrastructure of the internet that changes the social conditions of speech.

For example, there is a high risk of unintended consequences. Well-meaning content moderation systems have been abused to target racialized and other marginalized groups, algorithmic systems built on garbage inputs produce biased outputs, and so on. The answer is not more freedom of expression or more regulation, but rather something much more nuanced. For example, there should be no obligation to remove lawful but awful expression, as this would be an extraordinary interference with the right to freedom of expression. However, design features such as user empowerment tools to mute or curate content or presenting alternative news sources alongside flagged misinformation, may be justifiable, taking lessons perhaps from cases on time, place, and manner restrictions and advertising. Transparency obligations could include reporting on risk management measures taken to protect children and fundamental rights.

Second, there is a separate issue about whether these types of due diligence models even raise rights issues at all. A risk management model does not tell a company what to do, simply to do *something* and be accountable for that process. Therefore, it is unlikely that a court would view a social media company as undertaking governmental action and directly bound by the *Charter of Rights and Freedoms* for their public functions.

However, the decisions that flow from that risk assessment can implicate rights. The more specific the legislation, the greater the risk. For example, legislation that mandates demotion and promotion of content on recommender systems carries different rights risks. There is generally no right to an audience, and therefore demotion of content may be less problematic in law, although it is a grey area. However, mandating that content is amplified might violate the rights of the platform's freedom of expression as a form of compelled speech.

Similar issues are created by mandatory warnings/flags. Freedom of expression includes the right to say nothing or not to be forced to say certain things. The rights risk is greater in the US, where the compelled speech doctrine has been heavily litigated. In Canada, the proportionality analysis in s. 1 of the *Charter* weighs in favour of the constitutionality of warning labels, but it is always context-driven, and we have few cases to draw from (here and here).

There is good reason for the Government to start with a narrow list of illegal expression. If it follows in the footsteps of the EU, UK, and Australia, a new regulatory scheme will be born alongside all the growing pains that come along with it. But let's be clear: a vast swath of harms will remain unregulated, meaning we will continue to depend on corporate self-governance and selective transparency, if at all. For many types of harm, this is exactly as it should be if we are committed to freedom of expression. For others, the harms are too big to ignore. The next 10 years will be filled with debates about this penumbral space of expression, harms and algorithms.