Submission to Senate Legal and Constitutional Affairs Legislation Committee Inquiry into the Classification (Publications, Films and Computer Games) Amendment (Classification Tools and Other Measures) Bill 2014

Lyria Bennett Moses, David Vaile and Daniel Cater,
Cyberspace Law and Policy Community, UNSW Law[1]

UNSW Law's Cyberspace Law and Policy Community thanks the Legal and Constitutional Affairs Legislation Committee for the chance to respond to the inquiry into the Classification (Publications, Films and Computer Games) Amendment (Classification Tools and Other Measures) Bill 2014.

Introduction

Our main focus is on the Classification Tools regime in Schedule 1.

The content classification scheme, used as the basis for the Commonwealth's content censorship regime, originated in a time when most content was distributed in physical or electronic form from a centralised distributor or mass broadcaster. The Internet and other media resources have however created a fundamental challenge to the regime, and exposed it as potentially unsuitable to the contemporary market. This has led to a gradual erosion of the classification process for online material.[2] This Bill goes further in attempting to address the problem by replacing the existing reliance on Classification Board human judgment with "tools".

It may be intended to apply to small scale content decisions similar to those made at present, at least initially, but it seems also clearly designed to be capable of automated mass scale "classification" decisions. This is a functionality that may outsource the task from independent and objective professional classifiers to industries with a vested interest in certain outcomes. It may also at some later date be pressed into service in for instance a revived and expanded ISP-level mandatory content filter (the process for maintenance of the blacklist for which was one of its many flaws),[3] if or when a fully automated tool is ever approved and adopted.

Whatever the intended application, the ambition to do what is probably impossible -- converting a human judgment of many sophisticated cultural, ethical and developmental matters into a cheap and mass-scale automated "decision" -- may be a dramatic attempt to fit the current classification regime to a task which it possibly cannot meet, and which may undermine the balance of matters required under the current regime.

Human input

Classification decisions involve, in their essence, a human judgement about the nature and intent of the content, its context, the spread and balance of community sentiment about such content in its context, and its suitability for people of different ages. For instance, some classifications refer to these judgements: Refused Classification and other higher classifications use the terms "offend against the standards of morality, decency and propriety generally accepted by reasonable adults to the extent that they should not be classified", or more commonly, "in a way that is likely to cause offence to a reasonable adult".[4]

More generally, the *Classification (Publications, Films and Computer Games) Act 1995* (Cth) s 11 sets out a range of competing matters to be taken into account in classifying the various works. [5] The *National Classification Code* 2005 (Cth), cl 1. requires classification decisions to give effect, to the degree possible, to a range of potentially conflicting principles.[6] Under the legislation all of these matters must be taken into account in classification decisions, on a case by case basis, by a trained and professional independent classifier at the Classification Board.

This is the nature of the negotiated and publicly accepted basis for classification and censorship in Australia. The Classification Board, using skilled and impartial expert trained classifiers, uses their skill and judgement to apply these criteria for each classification. This allows for the wide range of different types and contexts of works, and ambiguities around unpopular or controversial material, while enabling in particular parents to draw useful guidance from these assessments.

The Classification Tools model appears potentially inimical to this approach, depending on what methods it uses.

Schedule 1 - Classification Tools

Recommendation 1

There should be a preliminary inquiry, before passage of the Bill, that investigates the nature of all process models that would be permitted to be used in operating a "Classification Tool".

The nature of the "Classification Tools", and how they would change the workflow and human input into the classification process is not clear from the Bill, which is mainly procedural. The Minister has wide discretion to approve any tool. At worst, there would be no human judgment applied to the necessary human judgment matters central to the classification process. A Google bot might do it.

There is also no indication how the tools would operate, the nature, inputs or rules of the processing algorithms that would give an "automated classification." At worst they would apply crude "machine learning algorithms" on image content criteria which fall short of exercising the judgement required by the Act.

Because there are many implementations of a tool that produces classifications that do not and cannot conduct the human judgment required by the Act, it should not be assumed that the formality of approving a tool and generating decisions from it meets the statutory and social intent of the Act. This should be established by investigation and demonstration, and be subject to evidence-based analysis. Creating a "black box" to make the sort of judgements required by the Act is too big a step to take without this level of fundamental and transparent inquiry prior to consideration of the Bill.

While some uses and mechanisms for creating and operating classification tools may be acceptable and fit within the ambit of the Act, there is nothing before us that constrains tools to do this, and no indication if there is anything other than the minister's discretion to constrain tools to within these limits. We believe this discretion is too broad.

Recommendation 2

There should be a requirement, if put into operation, that all approved classification tools are *transparent* in their operation.

The proposed Bill leaves open the definition of "approved classification tool", with the only limits in s 22CA(5). The Explanatory Memorandum (p 14) states:

> "An approved classification tool may take the form of a questionnaire and could be on a website or in the form of a computer program or other interface which allows a person to provide information regarding the relevant material. In response, the tool will deliver a classification decision and consumer advice for the relevant material"

The wide-ranging potential of the notion of "approved classification tool" ensures its applicability to future, yet to be conceived, tools. Such technology-neutrality is desirable in this context.

However, there are technology-neutral limits that could be put in place to ensure that tools can be tools operate *transparently*. In particular, in the case of algorithms that make binding decisions, transparency is essential to ensuring accuracy, legitimacy and accountability. Where a classification tool relies on a questionnaire in making decisions, it is important that the Regulator and the public understand how particular answers correspond to particular classifications. Where a classification tool employs a computer program to evaluate content, the source code should be made publicly available. In the event tools were to be designed based on machine learning techniques, third party evaluation would require at a minimum access to both the training data and the algorithm employed.

Thus while different types of disclosures will be useful for different types of tools, a technology-neutral requirement that tools be made transparent to the Regulator and the public ensures that it is possible to check the accuracy of tools with reference to the classification criteria. Transparency in classification systems was also raised by the ALRC.[7] A transparency requirement would both increase public confidence in the appropriateness of using particular tools in particular contexts and enable the identification of flaws and security vulnerabilities.

A requirement for transparency could form part of the written guidelines employed by the Minister (see s 22CA(4)) or form part of the list of requirements in s 22CA(5).

Recommendation 3

The Act should include, if the tools are to be used, specific mechanisms to ensure the appropriate use of approved classification tools:

It is essential that approved classification tools operate in a manner that is fair and transparent to those whose works are classified, those who have hosted content that is classified and the general public. The legislation contemplates the use of approved classification tools by people other than the creators, distributors and host of content. There is a risk in these circumstances that content will be classified based on false or inaccurate information fed into a classification tool by an unaccountable third party. As a result, the legislation would be enhanced by considering the following measures:

- There ought to be a provision that, where content is classified by an approved classification tool, the content creator, distributor and/or host (to the extent their identities and contact information are ascertainable) are notified of such classification;
- Similarly, any application to review the classification of content ought to be notified to the content creator, distributor and/or host (to the extent their identities and contact information are ascertainable);

- There ought to be a *requirement* for the Board to review a tool-based decision on receipt of an application, ideally within a set time limit, rather than the opportunity provided by s 22CH.

In order to be in the best position to make decisions about the ongoing usefulness of approved classification tools, addition provisions would be required:

- The Regulator should maintain a database to keep track of which tool or industry classifier classified particular material;
- The database should also be used to keep track of complaints about classifications, linking these to both the content concerned and the tool or classifier responsible. This will enable identification of tools or individuals who are frequently misclassifying content.
- Sections similar to ss 17B and 17C should apply to all types of content and all means of classification, including approved classification tools and, where relevant, the person inputting information into a classification tool that requires this.
- The list of approved tools should be subject to periodical review by the Minister to ensure that the tools used are up to date with industry best practice and current community standards. Outdated tools should be updated or removed from approval.

Recommendation 4
Ambiguity where classification tool loses its approval

Where a classification tool loses its approval, there are important questions concerning the impact on material already classified with that tool, in particular:

- Whether the public ought to be informed;
- Whether the producer or host of the material ought to be informed;
- Whether the person who classified the material using the tool ought to be informed;
- Whether the material ought to be reclassified.

None of these issues are dealt with in the Bill. Consideration should be given to the enactment of specific provisions to deal with this situation, providing answers to the questions raised above.

Schedule 2 – Referral of material to law enforcement agencies

Recommendation 5
The ALRC's recommendation in relation to limiting the scope of the prohibited category, and the implications that child abuse material is the target of law enforcement referral, should be implemented, and general referral of the current "Refused Classification" category not put into Schedule 2.

In its report, the ALRC proposed eliminating the current Refused Classification category and recognising a narrower Prohibited category.

The Explanatory Memorandum's justification for Schedule 2 revolves around content such as child abuse material. However, the same justification does not apply to the wider category of Refused Classification. The prompt law enforcement responses that are desirable and appropriate in relation to child abuse material will not be appropriate to all material that would be classified Refused Classification.

Schedule 3 - Exemptions

Recommendation 6

Given the purpose of the exclusion, the definition of social science should include a longer list than that proposed, or be left to its ordinary definition.

In proposed s 5C, "social sciences" is defined as economics, geography, anthropology, linguistics and other fields specified by instrument. The definition is said to be necessary in the EM to exclude "fringe elements, such as the occult or astrology". However, social science includes a diverse array of non-fringe subjects beyond the list specified in the legislation, such as history and law**.**

---

[1] See <http://cyberlawcentre.org/> The views are those only of the authors.

[2] *Broadcasting Services Act* 1992 (Cth) Schedule 5 envisages both online "Prohibited Content" in cl 20, as a result of classification decisions by the Classification Board, and "potentially Prohibited Content" in cl 21 arising from decisions by others guessing what the Classification Board would do. This "Prohibited" category was extended by a ministerial order in 2007 to add certain MA15+ and R18+ material to the prior X18+ and RC elements.

[3] See for instance David Vaile, Renée Watt, "Inspecting the despicable, assessing the unacceptable: Prohibited packets and the Great Firewall of Canberra", *Telecommunications Journal of Australia*, Vol 59 No 2 (2009) and the annotated bibliography also published in that edition.

[4] *National Classification Code* 2005 (Cth), cl 2(1) a-b.

[5] The matters are:

(a) the standards of morality, decency and propriety generally accepted by reasonable adults;

(b) the literary, artistic or educational merit (if any) of the publication, film or computer game;

(c) the general character of the publication, film or computer game, including whether it is of a medical, legal or scientific character;

(d) the persons or class of persons to or amongst whom it is published or is intended or likely to be published.

[6] The principles are:

(a) adults should be able to read, hear, see and play what they want;

(b) minors should be protected from material likely to harm or disturb them;

(c) everyone should be protected from exposure to unsolicited material that they find offensive;

(d) the need to take account of community concerns about:

(i) depictions that condone or incite violence, particularly sexual violence; and

(ii) the portrayal of persons in a demeaning manner.

[7] ALRC, *Classification—Content Regulation and Convergent Media* (ALRC Report 118) p27, [7.101], tabled 1 March 2012, at: http://www.alrc.gov.au/publications/classification-content-regulation-and-convergent-media-alrc-report-118