

eModeration

Contact information:

Tamara Littleton, CEO

T: 44 (0)20 7099 5310

M: 44 (0) 7776 138 642

E: tamara@emoderation.com

W: www.emoderation.com

B: <http://blog.emoderation.com>

eModeration Limited

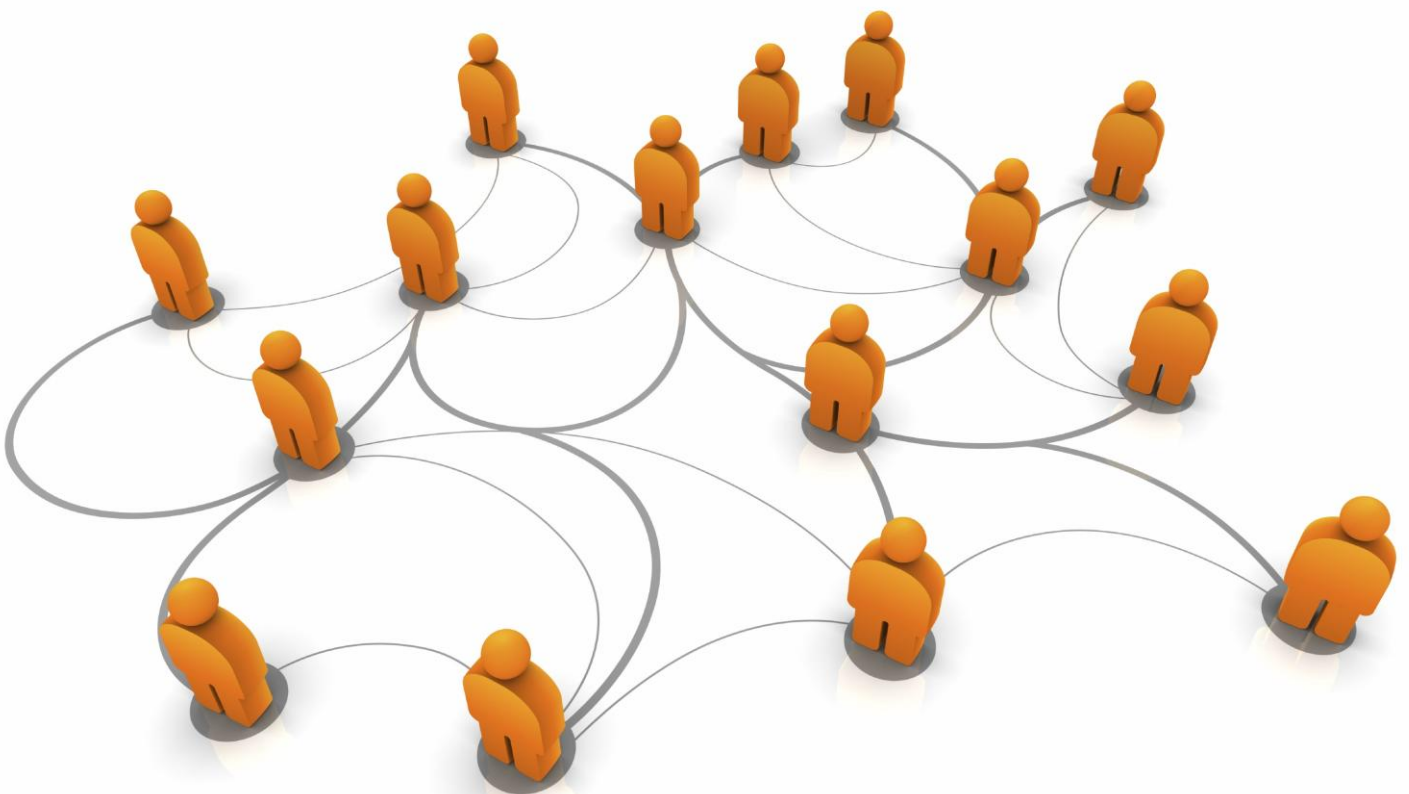
1 Cromwell Road

London

E17 9JN

White paper: Moderation in Social Networks

Date: 20th May 2010 (v3)



Moderation in Social Networks

A guide for brands and their agencies on moderating user generated content on social networks.

Introduction

With the recent announcement by Coca-Cola that it intends to “fish where the fish are” and that Coca-Cola and Unilever are shifting their digital focus away from traditional campaign sites and towards community platforms such as Facebook and YouTube, it is becoming increasingly the case that social media campaigns for consumer brands will usually include an element of social networking – whether that’s building their own networking platform, or creating profiles, pages, groups, competitions and other digital content within existing social networks. The nature of social networks means that these campaigns will include content generated by users: uploaded photos, comment streams, or videos.

Of course, social networks (and we’ll focus here on the ‘big four’: MySpace, Facebook and YouTube and Bebo, soon to be sold off by AOL) were not primarily designed as marketing tools. So the rules that govern them are not always clear to brands. In this paper, we will answer some of the main questions brands have around content on a social network site, covering:

- where their responsibility begins and ends on each of the big four networks
- what they should consider when setting up sites or pages on these networks
- what they should do to protect both audiences and the brand’s reputation from damaging content

Note that we are focusing on pages created by the brand as part of a social media campaign; we are not concerned here with defamatory content against the brand posted on third-party sites. Note also that the status of and law concerning social networks is changing very rapidly. We therefore aim to keep this white paper updated as far as possible, and would greatly appreciate any updates, corrections or comments.

The status quo: is content already safe on social networks?

Many brands wrongly believe that the content that will appear on their pages within a third party social network (particularly the big four of Bebo, Facebook, MySpace and YouTube) will be ‘safe’, as the networks have a legal responsibility to check content for illegal or abusive content. Not so. Social networks do not accept legal liability for everything on their sites – it simply would not be practical for them to do so.

In October 2009, a defamation suit brought by a teenager against Facebook was dismissed by a New York County Judge. The media website [Media Post reports](#) that fellow students of

the teenager, Denise Finkel, "allegedly posted false remarks indicating that Finkel "was a woman of dubious morals, dubious sexual character" and that she "engaged in bestiality," was "an IV drug user" and had contracted AIDS." The site reports that:

"The case against Facebook was widely viewed as unlikely to succeed because of the federal Communications Decency Act. Eric Goldman, director of the High Tech Law Institute at Santa Clara University, said back in March that Facebook was "categorically immune" from defamation lawsuits...

It goes on to say:

"Facebook praised the ruling in a statement: "We're pleased the courts continue to confirm that just as phone companies, email providers and postal services are not liable for misuse of their services to send inappropriate content, neither is Facebook," the company said. The company also said that it intends to seek sanctions in future lawsuits based on user-generated content. "These cases are dismissed time and again," the company said. "We think they should stop and that there should be penalties for wasting the court's time."

However, as a result of the suit, Facebook has made it clear within its [terms](#) that users own their own content.

The implications for brands are obvious. The ruling means that under US law, Facebook is not responsible for user-generated defamatory content against the brand on users' own pages, nor for illegal or abusive content posted on branded sites. So if you're creating a branded page on Facebook, and you don't want any illegal, spammy or abusive content posted, sort it out yourself. The social network is not responsible.

Clear? Well, almost. A case brought against Google by Italian law enforcement officers claims that Google's executives are responsible for a video (posted on Google Video in September 2006) showing a teenager with Down's Syndrome being bullied by a group of students in Turin. The clip was posted by one of the students involved, and was taken down by Google in early November, after complaints were made. Peter Fleischer, one of the Google execs named in the case, says in [his blog](#):

"It should be obvious, but none of us Google employees had any involvement with the uploaded video. None of us produced, uploaded or reviewed it." He also says that "Google complied with law enforcement requests to help identify the bullies, who were subsequently punished."

(Fleischer lays out his research into European laws on the responsibility of hosting sites [here](#).)

On 24th February 2010, Italian courts absolved the three of defamation but [convicted them of privacy violations](#). Google plans to appeal against the decision, saying that it is unrealistic to expect them to be able to moderate every piece of content. If the law holds, they doubt whether many internet platforms would be able to continue.

The argument seems to rest on whether social networks are content hosting sites, in which case they have no responsibility for content (in the way that a telecoms provider wouldn't be responsible for a nuisance call), or media sites that produce content – in which case they **do** have a responsibility (in the UK, the BBC lays out its assumption of responsibility for content on its own sites [here](#)).

In Italy, as the case against Google continues, the definitions are based on the European e-Commerce Directive, brought into law in 2000 – which clearly wasn't designed for Web 2.0, social networks and the like. In the US, a combination of the Communications Decency Act (CDA) – which was originally designed to provide protection from defamation claims - and the Digital Millennium Copyright Act (DMCA) insulates providers of online content from potential liability. Specifically, under the CDA, providers of 'interactive computer services' (such as social networks) are protected from liability for information submitted by a third party or consumer. This provides a certain 'safe harbour' for social networks, which means they don't have to accept liability for content. However, there are some caveats to this: a site can be held accountable for deliberate trademark or intellectual property infringement in the US, for example.

It is important to note that there are no international binding laws governing how site owners should behave. Which country should claim jurisdiction over a particular site is often the first issue considered in a litigation case. Of course, a site hosted in the US is subject to US law – but where jurisdiction becomes less clear is, for example, when a case is brought against an international site targeting consumers outside of its hosting country (as in the Google / Italy case). If a US-hosted site is actively targeting people in another country, it is likely to still be held to account under that country's laws. So, broadly, if a network targets people in France, it will be accountable under French law. There are often issues of laws conflicting. For example, US and UK data protection laws are very different (the US is much less regulated than the UK). It is unclear whether UK data stored in the US should be governed by UK or US data protection laws.

Both US and UK law appears (from test cases in both countries) to apply 'safe harbour' from liability where all reasonable measures have been taken to prevent infringement of, for example, copyright. If a network is seen to be actively encouraging illegal or offensive behaviour, then it would lose its 'safe harbour' status. So if networks have active policies in place to prevent rule-breaking where possible, they are seen to be taking reasonable care to protect brands and users. In the EU, an Information Society Service Providers (ISSP) is not liable if : a) it does not have actual knowledge of the unlawful activity or information or upon obtaining such knowledge or awareness and it acts expeditiously to remove or disable access to the information; and b) the user was not acting under the authority or the control of the ISSP.

So, if a defamatory comment is uploaded onto a social network page, the social network is not expected to know they're there. But once a user has reported that comment to the network, it has a legal responsibility to deal with it (as in the Godfrey vs Demon case in the

US). Turning a blind eye to inappropriate content is not enough. It is now generally accepted that a network that accepts user-generated content should have clear terms that allows it to edit or take down material posted by users. This shows that the network is exercising 'reasonable care' to avoid offensive, illegal or defamatory content.

Note that this paper should not be taken as providing a definitive legal framework for online content (we leave that to the lawyers). The law is still unclear, and test cases are fast changing the interpretation of outdated laws. But what it does show is that brands shouldn't assume that operating within the boundaries of a social network (creating a branded fan page, for example) gives them immunity in law. They too, should assume responsibility for taking reasonable care to protect themselves and their users.

Who is responsible for keeping users safe?

As we've seen, the legal responsibilities are unclear. But media owners and social networks are starting to tackle their moral duty to keep their users safe. In February 2009, 17 web firms and social networks (including all of the 'big four') signed an agreement in Luxembourg to protect under-18s using their sites, through a number of practical steps including (as reported on the European Union's portal, [Europa](#)):

- *Providing an easy to use and accessible **"report abuse" button**, allowing users to report inappropriate contact from or conduct by another user with one click*
- *Making sure that the full online profiles and contact lists of website users who are registered as under 18s are **set to "private" by default**. This will make it harder for people with bad intentions to get in touch with the young person*
- *Ensuring that **private profiles** of users under the age of 18 are **not searchable** (on the websites or via search engines)*
- *Guaranteeing that **privacy options** are **prominent and accessible** at all times, so that users can easily work out if just their friends, or the entire world, can see what they post online ¹*
- *Preventing **under-age users** from using their services: if a social networking site targets teenagers over 13, it should be difficult for people below that age to register*

¹ Facebook's recent [privacy changes](#) in December 2009 caused much controversy in requiring users to reset the 'improved' privacy settings. The default settings in the transition tool suggest that most users make many of their updates, including the posts they create and their "about me" information, available to 'everyone'. Users will no longer be able to restrict access to some basic information: gender, what pages they are fans of, and will also have less control over what information about them is shared via the Facebook API. Following this, Facebook have further antagonised those concerned about privacy with the introduction of ['instant personalisation'](#) and [Community Pages](#) in March/April 2010 (more about Community Pages later in this white paper). At the time of the last update, Facebook are considering [simplifying their privacy settings](#).

The Social Networking Task Force, who brokered the agreement, met on Safer Internet Day in Feb 2010 to assess what actions have been taken - the results are published [here](#).

The UK government is taking steps to ensure that [children are educated](#) on the potential dangers of the internet, and the UK Council for Child Internet Safety (UKCCIS) reported in December 2009 its initial strategy recommendations that websites install a panic button to report inappropriate behaviour. According to the [Telegraph](#), this was being considered by 140 companies - including Bebo, who have now [introduced a 'report abuse'](#) button linking to [CEOP](#), so that children can report abusive or grooming behaviour, easily and quickly.

Subsequently, in early 2010, CEOP and Facebook were involved in [a very public argument](#) about whether such a 'panic button' was appropriate for the Facebook site, which eventually appeared to be resolved (for the moment at least) with the introduction of a new [Safety Centre](#) and a promise of free space on the site to be given to safety organisations.

YouTube, in turn, released a 'Safety Control', a setting to filter out potentially objectionable content on Feb 10th 2010. See our [blog post](#) on the subject for more information.

Each network has a feedback mechanism which allows users to report content that is inappropriate (although YouTube allows comments only to be flagged as 'spam' rather than the 'inappropriate' flag, which is reserved for videos only); it is then vetted by teams of people who trawl through content that has been flagged up by users and delete material that goes against the terms of the site (for example anything illegal, abusive or pornographic). [Newsweek profiles](#) the teams at Facebook and Google, reporting the 'informal' rules that are drawn up to decide what goes and what stays.

These are mostly positive moves, but they are primarily designed to help social network users (children in particular) report abuse, not to prevent inappropriate content from being posted in the first place.

The BBC guidelines to editorial staff make the point that the social networks keep information on how they enforce their terms close to their chests:

“One problem is that while social networking sites may publish clear rules of acceptable behaviour for their users, they are often very reluctant to share much information about how they intervene or to what level.”

The age issue is the elephant in the room. On each network, users are supposed to be 13 or above. However, it is fairly commonly accepted that many children under 13 use social networks – a very hard thing for the networks themselves to police. In May 2010, [a survey](#) of 1000 8-15 year old girls said that Facebook was one of the most important parts of their lives, clearly indicating that the site was being used by underage children. As part of the February 2009 Luxembourg agreement, all the main networks use cookies to prevent users from changing their age (i.e. if they register as 12, the site will reject them; if they sign in again

as 13, the site will remember their previous registration using that computer, for as long as the cookies are not cleared), but this is obviously not effective enough. In the US, sites are held responsible under the Children's Online Privacy Protection Act (COPPA) that prevents a network from actively or knowingly collecting personally identifiable information from a child under 13 without the explicit permission of parents or guardians. The truth is, children will always find ways round age filters, and brands should assume that younger children could be among those using the social network on which they have a presence.

What is the risk for a brand?

There are obvious risks for brands which don't moderate the content posted onto their own branded channels on social networks (i.e. on a branded fan page, company profile page, group or channel). The most important is the safety of their users, particularly brands which are marketing to children or teenagers. The importance of providing a safe environment for children goes without saying, and brands have a duty to ensure that children are not exposed to abuse, bullying or even illegal content posted by unscrupulous users of their social network pages.

There is also a reputation risk. Like it or not, content posted on a branded page will be associated with that brand. No responsible company wants to be associated with bullying or inappropriate content on their social network pages. On a practical level, users won't come back to a site that is rendered unusable by people posting comment spam, or irrelevant messages.

Who makes the rules on social networks?

Each of the big four social networks has its own user policy and terms and conditions – which we'll examine in more detail, below. But brands can state their own rules within their branded pages. What a brand considers acceptable may differ from what the network considers acceptable (the use of swearing, for example, or nudity, or coded messages that might be offensive or give away personally identifiable information). If in doubt, set your own rules.

Can a brand moderate content that is held on a social network?

It is possible for brands to moderate the content that is uploaded to their pages on social networks. The rules for each are different, so we've laid out some of the key points for each of the 'big four', below.

1. [YouTube](#)

- a. *What's possible:* pre- and post- moderation of content are both possible, and are very easy to set up. Most user-generated content can be sent to a designated inbox before it goes live on the site, so can be pre-moderated. All comments can be set to be pre-moderated. This includes all comments on videos, channels and anywhere else. Comments can even be turned off or set to allow "only friends" comment. However, you may only moderate comments on UGC that the brand itself has uploaded – see section c. below.

- b. *What you can pre- or post- moderate:* friend requests, video responses, comments (on brand generated content) and profile posts. There is no way to post moderate friend invitations on YouTube. They are all pre-moderated. The rest can be set either way.
- c. *What you can't pre- or post- moderate:* avatars of friends. But you can reject friends with offensive images for avatars or offensive names and the issue with avatars on YouTube only comes into play on channel comments, where there is a thumbnail image by the post. Otherwise no avatars appear on the comments. To be noted though, is that, the moderation of comments on video responses belongs to whoever has uploaded the video response on their page. There will be a link to the video response from the brand page, and the only way to disassociate the brand from any undesirable content in this way is to delete the video response.
- d. *Issues:* users can change avatar images quickly and without flagging it up, so it would be almost impossible to track. Moderation of comments on response videos (see above) – although there would seem to be a black list filter available which would block 'objectionable words' from being uploaded.

2. [MySpace](#)

- a. *What's possible:* pre and post moderation are more difficult to do, but are still possible. MySpace is very responsive to moderation. Comments, messages and friend requests come to a central inbox for pre moderation (this is a significant difference from Facebook's process); or it can be post moderated
- b. *What you can pre- or post- moderate:* fan requests, comments, videos, images, blogs (and their comments) and private messages
- c. *What you can't pre- or post- moderate:* fan updates (live feed) that roll on the front page. However, you can remove these from the profile page altogether by blocking the fan concerned.

3. [Bebo](#)

- a. *What's possible:* Bebo's moderation process is very similar to MySpace. Content comes to an inbox where it can be pre moderated, or content can be post moderated
- b. *What you can pre- or post- moderate::* fan and friend requests, and comments
- c. *What you can't pre- or post- moderate* live feed (as with MySpace). The live feed includes comments on fans' profiles and can't be changed, although you can prevent them from appearing by removing the friend's profile. .

- d. *Issues:* Bebo doesn't check usernames, so moderators need to block unsuitable names (for example, if you have a user named 'sexybiatch' on a site aimed at 13-year olds)

4. [Facebook](#)

- a. *What's possible:* Facebook does not offer pre- moderation, so the only option is to post-moderate and delete comments that violate site rules, or to block persistent offenders
- b. *What you can post-moderate:* wall, photos, discussions, notes and comments
- c. *What you can't post- moderate:* Reviews can't be deleted: only reported through to Facebook. It appears hard to remove 'Graffiti' also
- d. *Issues:* Facebook is very difficult to moderate. It has not been set up to make it easy for brands to moderate content; as a result there are various companies developing moderation tools as well as apps specifically for Facebook.

Facebook have now launched their 'Preferred Developer Consultant Programme' to connect people to the resources they need to build with Facebook products and technologies, (especially important now that brand competitions need prior approval from Facebook). Of these, [Context Optional](#) and [The iPlatform](#) have recently developed fairly robust moderation solutions for Facebook, although limited to wall content (i.e. not third part apps).

Recently Facebook started to beta a crowd-sourcing idea to deal with its moderation issues. The 'Facebook Community Council' (currently being tested on a small number of users) is a way for users to decide whether reported content violates Facebook policies or not. Facebook said: ""We've found that people aren't shy about reporting content they come across that looks suspicious, and this is just another way of leveraging the Facebook community to help maintain the site's trusted environment. It's still in an experimental stage, and we're currently testing the application with only a very small number of users."

[Community Pages](#). Recently introduced by Facebook in order to give 'fans' a place to talk about non-brand topics, because they aggregate all mentions of the topic (or brand) onto the page, the Community Pages pose a severe [reputational threat](#) about which a brand can do nothing as they have no editorial control over the page.

Should brands moderate content on a third-party site?

How well do you know your audience?

We are often asked by brands whether they should moderate content that is held on a third-party site, such as a social network. Some brands still believe in their audiences enough to assume they don't need to filter the content they'll upload. But the nature of social networks means that they are inherently open to public content (and therefore public abuse) – not just trusted 'fans' of a brand.

However, we do see different behaviour on, for example, gaming sites (where the audience tends to be older, male, and finds swearing more acceptable, for example) than on sites for younger audiences. But however much a brand trusts its target audience, you simply can't guarantee that someone with bad intentions won't infiltrate a site aimed at a younger audience; or that younger viewers won't enter an open site on a social network, even though the brand's audience is usually older.

Negative associations for the brand

The main concern is the negative associations that a brand might suffer as a result of inappropriate content being posted on its page. Many users will assume that brands check the content that goes onto their pages and so, if (for example) racist comments were to appear on a YouTube channel, users might assume the brand endorses those comments. Evidence of how strongly brands view this impact is demonstrated by Tesco's stance in a reversal situation: in May 2009, [Tesco pulled its advertising](#) from non-Tesco Facebook pages, after its ads were seen next to Holocaust denial and BNP pages. Other major advertisers quickly followed suit. Tesco now will only advertise against content it controls and in a [review of Q4 2009](#), AdSafe found that a large portion of user generated content is a significant concern to advertisers with high sensitivity brands or high sensitivity industries.

Protecting users

This is much-discussed elsewhere, but is probably the most important reason for moderation. Brands which create content on a social network have a responsibility to protect their users from exposure to illegal or inappropriate content, bullying by other group members - and importantly, from themselves – from sharing personally identifiable information that could be used to target vulnerable groups such as children.

Can brands stop people saying negative things about them?

Moderation is not censorship. We believe absolutely that if you are using social media to engage with consumers, you should listen to those consumers – whether what they have to say is good, or bad. Social media is about listening and engaging, and negative feedback should not be censored because the brand doesn't want to hear negative things about itself. Apart from anything else, consumers do not respond well to censorship, and you could be creating an even bigger problem.

What should a brand look for when moderating content?

The obvious issues to avoid are bullying, abuse and illegal content. But there are other, possibly less obvious things that brands should look out for. Some examples that we've come across include:

Users with names that include abusive or obscene words. [Starbucks faced this problem](#) when a user had included a swastika in their profile picture. Options to tackle this are: block the user outright; or contact them and ask them to change their avatar. If they refuse, or they change it back, block them (sometimes it may be necessary to involve the social network in this action, although it should always be possible).

Obviously off-topic posts. If a user posts something that is obviously off-topic, it should be considered to be spam (particularly if it includes a link – this could take the user to a website that is infected with malware or contains offensive images, for example). Spam will disengage users and make your site less relevant and interesting to them. The solution? Don't publish the post, or delete it. If necessary, block the user.

Non-fans. By 'non-fans' we mean people who are leaving harassing messages (threatening 'chain mail' style messages, for example) or people who are just trying to sell fans a product. Options to tackle this are: block the post and if appropriate, contact the user to explain why; or block the user if this is possible. Note that in an open group on Facebook, you can't delete a user's profile, but persistent offenders can be reported to Facebook (and on other sites, through the relevant 'report it' channel).

In summary

To sum up, the risks of not moderating a social network page, group or channel are the same as not moderating a brand-created website or online environment. Moderation is particularly important where brands are likely to attract younger audiences, but those that attract adults should be aware that they will be open to younger audiences by default, within a social network. These are not closed communities.

There is little, if any, consistency among the big networks around how they are run or moderated; and no real legal framework, which is much-needed to provide clarity around who is responsible for what. This is not a new debate – it has raged around illegal web content (and whether the ISPs are responsible) for years. Social networks set their own policies, and their own moral codes. Brands should be aware of this, and take responsibility for their own content on social networks, rather than rely on the networks to do their work for them.

Resources / further reading

- [BBC Editorial Guidelines on Using Social Networks](#)
- Jason Falls on [Facebook Group And Brand Page Best Practices](#) (from 2008, but still good points to make)

- [Social Networks Websites Review \(Bebo\)](#)
- [Social networks Websites Review \(Facebook\)](#)
- [TaylorWessing's guide to Liability for User-Generated Content](#)
- [ReedSmith: Network Interference: A Legal Guide to the Commercial Risks and Rewards of the Social Media Phenomenon](#)
- [Mashable's Facebook GuideBook](#)

Please feel free to get in touch with us and suggest other resources or updates to this white paper.

About eModeration

Founded in 2002, eModeration Limited is an international, specialist user-generated content moderation company. It provides multilingual community management and content moderation to clients in the entertainment and digital publishing industry and major corporate clients hosting online communities and consumer-driven ad campaigns.

eModeration's team of moderators and staff are the key to eModeration's success and excellent client list. eModeration draws on the expertise of carefully recruited and trained moderators located mainly in the US and Europe with specialist editorial and community moderation skills, which are matched uniquely to the client. The company can moderate 24/7 and provides cover for over 40 languages. All its moderators are managed online from eModeration's headquarters in London, United Kingdom.

For further press information, or to speak to Tamara Littleton, CEO of eModeration, please contact:

Kate Hartley
Carrot Communications
Tel: +44 (0)771 406 5233
E: emoderation@carrotcomms.co.uk
Twitter : @katehartley

© eModeration Limited 2010

This document is the intellectual property of eModeration Limited and may not be duplicated or disclosed to any third party without the written permission of an authorised officer of the company.