Regulating Behavior in Online Communities

Sara Kiesler, Robert Kraut, Paul Resnick & Aniket Kittur,

Introduction

"One bad apple spoils the barrel"

In thriving communities, a rough consensus eventually emerges about the range of behaviors the managers and most members consider acceptable, what we will call normative behaviors, and another range of behaviors that are beyond the pale. *A Rape In Cyberspace*, the newspaper report by Julian Dibbell (1993), describes a classic example of unacceptable behavior in LamdaMoo, an early virtual environment. Mr. Bungle, an avatar in the online community, wrote a program that forced two avatars controlled by other participants to have virtual sex with him and with each other, and to do brutal things to their own bodies. In describing the event online the next day, one of the victims begged ""I am requesting that Mr. Bungle be toaded for raping Starsinger and I [stet]," where "toad" is the command that would turn Bungle's avatar into a toad, annihilating the character's original description and attributes. Within 24 hours, 50 other characters also called for his toading. Three days later the community had a real-time discussion of the issue. An system administrator who observed this discussion eventually ran the toad command to eliminate the Mr. Bungle character. Although LamdaMoo did not have a policy against cyberrape, when one occurred in its midst, this action instigated widespread discussion and crystallized a view among many inhabitants of what were correct and incorrect types of behavior in this community.

Communities differ on what behaviors are normative and which are not. Personal insults may be the primary way to interact in one community, but frowned upon in another. Wikipedia expects writers to adopt a neutral point of view when writing articles, while the Huffington Post expects

guest bloggers to express a viewpoint. PsychCentral.com, a site with 160 health support communities, forbids members or outsiders from conducting any type of research on the site for publication or educational purposes (PsychCentral, 2008). JoBlo's Movie Club wants only on-topic posts on it forums. Its list of rules emphatically states, 'Our board is for MOVIE TALK only…This is … not a place for you to discuss your personal life or boohoo about how your lover just broke up with you. (Joeblo Movie Club, 2005)" As we will explore in this chapter, the normative behaviors may be codified and articulated or may be left implicit, and they may be contested by some members at some times, but most of the time, most people will agree about behaviors that are acceptable and those that are not.

Having a rough consensus about normative behaviors can help the community to achieve its mission. In many technical and health support communities, it is expected that responses to questions will be supportive rather than antagonistic, aiding the mission of helping the members deal with problems they are having. Many open source software development communities expect that discussion of plans, features and bugs will occur in open discussion lists, rather than in private email conversations between developers. The Apache Web Server project notes, "Public forums, which include all developer and user mailing lists, wikis, and issue reporting systems, are essential to The Apache Software Foundation (ASF). We strive to do our work in public forums in a spirit of transparency and openness. They are our preferred means of communication." This norm is functional in groups where developers are not collocated and work is interdependent. It enables project members to maintain an awareness of the state of the project, by making available information that can affect the work of others. The neutral point of view norm in Wikipedia supports the community's goal to write a trustworthy encyclopedia, while the norm that editors take particular care when adding information about living persons can reduce threats from lawsuit.

Expectations about how to handle conflicts are especially important to keeping a community productive. Conflicts are inevitable in social interactions, but if they become personal and escalate, they can derail a community, taking attention from its mission. Extended "flame wars" can start between two people and grow to involve many. Before they end, some people may be sufficiently alienated to leave the community. The LinuxChix community, which encourages women's participation in Linux open source software development, prides itself on its nurturing atmosphere, unlike other Linux communities "dominated by flame wars and ego battles (Vesperman & Henson, 2004, p. 1), which tend to drive people away." Many communities have behavioral norms governing conflicts, such as avoiding personal attacks, moving conflicts to special locations, or using special mediation processes. In Wikipedia, a common source of conflict is different views about what content should appear on a particular page. This can lead to "edit wars" where editors repeatedly undo each other's edits in an attempt to make their own preferred version of the article visible (Kittur, Suh, Pendleton, & Chi, 2007). These edit wars can even occur in seemingly non-polarized topics. For example, Viegas et al demonstrated an edit war over whether a kind of chocolate sculpture called "coulage" really existed and whether the paragraph describing it should appear on the article (Viegas, Wattenberg, & Dave, 2004). The existence of edit wars resulted in the three-revert rule in Wikipedia, which holds that editors may revert an article to a previous state a maximum of three times per day.

Not everyone will comply with the consensus standards of normative behavior all the time. The Internet is filled with "trolls" and "griefers," people who derive satisfaction from disrupting communities. Trolls pose as legitimate members and post inflammatory comments designed to provoke other members. For example, the website democraticunderground.com is a community for liberals to post and discuss news. There, a concern is trolls "who professes complete faith in the progressive cause, who deliberately works to destroy it by claiming falsely that our displays of courage and strength are actually a weakness." They post comments like, "I am a lifelong Democrat but I just feel the party is being damaged by association with Howard Dean/Russ Feingold/DailyKos (bunkerbuster1, 2006)" Both essays by game designers (Bartle, 1996) and empirical factor-analytic studies by social scientists (Seay & Kraut, 2007; Smith, 2007) indicate that some players in online games are motivated by causing problems for other players. For example, in World of Warcraft, the popular role playing game, a griefer might engage in what is known as "corpse camping", when the griefer remains near the corpse of other players after killing them in game and repeatedly re-killing them whenever they resurrect themselves.

Trolls can do a lot of damage. Consider, for example, the alt.hackers newsgroup. Ordinarily, to post to alt.hackers, newcomers must hack into the board. Once in, the poster is expected to include an ObHack—information about technology shared with others. Wysocki (2002) describes the regulatory breakdown in this community after a Usenet bulletin board in Italy started leaking messages into the alt.hackers site. A member of the alt.hackers complained about these new posts, written in Italian, "why, oh, why do these RUDE BASTARDS *still* post here in a language only morons would speak?" Another member replied that this post, besides being racist, had not included the obligatory ObHack. A spirited discussion and pursuit of the mysterious foreign messages ensued. An Italian hacker, Venom, bragged, "They can read any message you post, the complaints too, they simply don't care, and taunt you." Ultimately, the Italian hackers left, but at least one member of alt.hackers quit publicly in disgust because an alt.hacker member had violated the group's secret procedures, "Idiot. This whole situation was brought about by someone posting instructions on how to bypass the /one/ thing that prevented lusers . . ."

Another threat comes from manipulators. They do not gain utility from disrupting the community, but from getting the community to produce particular outcomes. For example, in a community like Yelp or TripAdvisor that reviews and recommends commercial establishments, manipulators may want to pump up the ratings of a particular venue. Or on Wikipedia, they may want particular pages to reflect their point of view rather than a neutral point of view. Manipulators will make use of multiple "shill" accounts to carry out their manipulations.

We refer to trolls and manipulators as outsiders because they have no vested interest in the community functioning well. It is especially difficult to deal with them because social sanctions (being disliked or publicly disparaged, or losing status in the community) may either have no effect, or, in the case of trolls, actually increase their utility.

Insiders, however, may also violate behavioral norms. Chapter 3 on the socialization of newcomers notes that new members of a community may often act non-normatively because they simply do not know the rules of the community. Failing to understand the norms of a community may also result from a number of causes besides members' lack of experience in the

community. For example, some members may have cognitive or social impairments that make it difficult for them to infer rules from observations. As Burke and her colleagues notes, young adults with Asperger's Syndrome or other disorders on the autism spectrum are socially awkward in part because of their difficulties in generalizing social norms from repeated exposure to examples of people following them (Burke, Kraut, & Williams, 2010). These authors describe young men who were cut off by communication partners after sending them a few hundred text messages over a two day period or because they revealed too much information about their childhood (i.e., a "creepy" amount) when trying to reconnect to grade-school friends on Facebook.

Even insiders who know the norms may not comply with them. Existing norms may be contested, and they may follow other expectations that they think *should* be the norms for the community. They may also fail to comply simply because it is in their own interest to do so in particular situations.

Social scientists use the term "social dilemma" to describe situations where everyone is better off if everyone complies with the norms than if no one does, but each individual is even better off if she does not comply while the others do. One form of social dilemma is called a public goods problem, where everyone is better off if everyone contributes some effort to the community but there is a temptation to free-ride on others' contributions. We consider ways to motivate public goods contributions in Chapter 5. Another form of social dilemma is the common pool resource or public bad problem, where individuals are tempted to take actions that use up or pollute a shared resource. Garrett Hardin, in his famous paper, Tragedy of the Commons (1968) explained the problem:

> *"Picture a pasture open to all. It is to be expected that each herdsman will try to keep as many cattle as possible on the commons. . . . As a rational being, each herdsman seeks to maximize his gain. Explicitly or implicitly, more or less consciously, he asks, 'What is the utility to me of adding one more animal to my herd?' This utility has one negative and one positive component. The positive component is a function of the increment of one animal. Since the herdsman receives all the proceeds from the sale of the additional animal, the positive utility is nearly +1. The negative component is a function of the additional overgrazing created by one more animal. Since, however, the effects of overgrazing are shared by all the herdsmen, the negative utility for any particular decision-making herdsman is only a fraction of -1.*
>
> *Adding together the component partial utilities, the rational herdsman concludes that the only sensible course for him to pursue is to add another animal to his herd. And another; and another. . . . But this is the conclusion reached by each and every rational herdsman sharing a commons. Therein is the tragedy. Each man is locked into a system that compels him to increase his herd without limit—in a world that is limited".*

In economics and psychology, different versions of this fundamental conflict have been modeled as games and tested in experiments. Walker, Gardner, and Ostrom (Walker, Gardner, & Ostrom, 1990) used this setup: each subject begins with an endowment of tokens. Each token "invested" in the common pool resource market earns 23 tokens minus a quarter of the total tokens

contributed by all the subjects. Investing in this market is analogous to grazing a sheep on the common pasture: it creates some value for the herdsman but reduces the value of every other sheep using the pasture. This situation sets up a social dilemma because it is in each subject's self-interest to put all their tokens into the market that uses the common pool resource. Collectively, however, this is a disaster, because they all earn less than they could have had they coordinated on a strategy of putting fewer tokens in (using less of the common pool resource).

In online communities, people's attention (or bandwidth, as Kollock & Smith, 1996 describe it) is a limited resource. People may be motivated to participate in a community for many reasons: to amuse themselves, to help a favorite cause, to utilize their expertise, to get people to talk to them, to enhance their reputations, or because they expect that they will receive useful help or information in return (Constant et al , Butler et al). These multiple motivations can lead many people to post messages and be actively involved, but if their contributions are trivial or silly, these contributions consume others' attention for little benefit. Many low quality contributions create a social dilemma wherein these contributions drown out the worthy contributions and exhaust the available attention. Like the herdsmen using up the limited grassland for his own herd, members of the community may use up everyone's attention on messages that meet their own needs but not those of the recipients.

Similar reasoning applies to the cheats that occur in many multi-player online games, which allow one player to gain advantage over other players, while polluting the experience for other players. For example, in online role player games, Godmodding is the term used when players create a character that is virtually indestructible. As one commenter discussing this practice for the Marvel Heroes RPG [Role Playing Game] notes, "this is frowned upon by other members of the RPG and is extremely annoying (47jamoo, 2007)."

Ostrom studied a number of institutions that have successfully self-governed common pool resources over a long-period of time (Ostrom, 1990, p. 90). The resources included forest and grazing grounds, fisheries, and water for irrigation. She identified seven design principles that seemed to underlie their success. We will revisit several through the chapter, including the need for: community participation in rule-making, monitoring, graduated sanctions, and conflict-resolution mechanisms.We adopt the term regulation to describe any efforts to decrease the frequency of non-normative behaviors or lessen their impacts on the community. Lessig (1999) identifies four elements that regulate behavior online: laws, norms, markets and technology (or code or architecture, as he called it). Laws are rules propagated and sometimes enforced by government, and are external to the community itself. For example, most western countries have laws against the creation, possession and distribution of child pornography, images involving the sexually explicit activities involving a child, and these laws are often enforced when these images are distributed over the Internet (Akdeniz, 2008). Lessig argued that laws are difficult to enforce on the Internet, and urged policy makers to consider other means of regulation, especially technology. In any case, online community designers will generally have little control over the laws governing their communities, though they can publicize them more or less and more or less proactively cooperate with law enforcement.

This chapter considers means of regulation that fall into the other three categories. Some design alternatives, like making norms more salient or shaming people who violate them, create

psychological motivations for compliance. Some, like reputation systems and internal currencies, create economic incentives. And some employ technical means like reversion tools, moderation systems to prevent and recover from bad behavior. Often, means of different kinds complement each other. For example, as we shall see, reputation systems function better in combination with technical mechanisms that limit the ability to create new accounts.

The chapter begins with an analysis of ways to limit the damage that bad behavior causes when it occurs. Next, we consider ways to limit the amount of bad behavior that a bad actor can do. Finally, the third and longest section consider ways to encourage compliance with norms through psychological and economic incentives.

Limiting Effects of Bad Behavior

In asynchronous conversation communities, posts can be screened. In some email lists, for example, a moderator has to approve each message before it is forwarded to all the members. In many forums, moderators can remove inappropriate messages after they are posted, or move them to other forums where they may be more appropriate. Messages may also be degraded, but left in place. For example, "disemvoweling" removes all the vowels from a message. Readers are quickly aware that the message has been degraded, but can still, with effort, read it. Alternatively, posts can be labeled or rated, and individual readers can sort or filter what they read. For example, at Slashdot, where comments are scored from -1 to +5, the default reading settings hide comments with scores of 0 or -1, but individual readers can change the settings.

All of these techniques limit the impact of inappropriate messages, because they reduce the number of people who will read them. To the extent that people are psychologically vested in the community and its reaction to their posts, these techniques also act as sanctions against the person. We will analyze how to make sanctioning mechanisms effective later in the chapter.

Design Claim 1: Moderation systems that pre-screen, degrade, label, move, or remove inappropriate messages limit the damage they cause.

One of the problems with moderation systems is that people may not agree with the moderators' decisions. If they do not accept the legitimacy of the action, they may take further actions (e.g., posting additional inflammatory messages). The net damage may even be greater than what the original message would have caused, had it gone unmoderated. Thus, moderation systems will be more effective when the decisions are perceived as more legitimate. A similar logic applies to the use of reversion tools.

Moderation actions that do not silence the speaker will be met with less resistance. Many communities try to keep conversation limited to designated topics. When off-topic conversation arises, if it is redirected to another, more appropriate forum, either by directly moving it or posting a response suggesting where the conversation should be continued, people are less likely to insist on their right to talk about the topic in its original location. One common approach is to have a special space or spaces where the normal rules of behavior do not apply. For example, an off-topic forum can handle the messages that don't belong elsewhere. Online community

organizer Caleb Clark, in an interview with Derek Powazek, reports that he has found it effective to create an "Outside" space for flames and fights: (Powzek, p.113)

> *"Well, I thought why not have an "Outside" in an online community? A place to go when what you are doing is bothering other people, but you still need to do it. Most people hate to be told what they can't do. But they don't seem to mind so much a little structure on what they can do. So it's worked great whenever I've tried it. When I encounter flames sparking up, I send an email saying, "Take it outside." It's a great re-director of bad energy in a community. Interestingly, it seems to take the gas out of flames very fast, since there are not a bunch of people 'watching' the flame."*

Design Claim 2: Redirecting inappropriate posts to other places will create less resistance than removing them.

One source of legitimacy comes from the notion of procedural justice, that sanctions are given through a fair procedure. In fact, people would rather take a more severe punishment after they have had their "day in court" than a milder punishment without any hearing (Tyler, 1990). In two-party conflicts, people prefer arbitration after both sides present information on their case over arbitrary top-down decision making (Ross and Conlon, 2000). People's perceptions that they have been treated fairly are greater if procedures (a) are applied consistently across people and time, (b) are free from bias, (c) collect and use accurate information in decision making, (d) have some mechanism to correct flawed or inaccurate decisions, (e) conform to personal or prevailing standards of ethics or morality, and (f) ensure that the opinions of various groups affected by the decision have been taken into account. (Colquitt, Conlon, Wesson, Porter, & Ng, 2001, p. 426; Leventhal, 1976). In conventional organizations, perception of procedural justice is associated with such desirable outcomes as satisfaction with the outcome of decision, evaluations of decision makers, organizational commitment, willingness to engage in voluntary, organizational citizenship behaviors and job performance (Colquitt, et al., 2001, Table 5).

This research suggests that online community member will be more satisfied with moderation decisions if they are delivered through fair procedures. Thus, legitimacy will be enhance if criteria for moderation are clearly spelled out and consistently applied. It will also be enhance if people have a chance to argue their cases with the moderator and even appeal to a third party. Of course, those procedures may be costly, taking time of moderators and other authorities, so they may not be practical in all situations

Design Claim 3: Consistently applied moderation criteria, a chance to argue one's case, and appeal procedures increase the legitimacy and thus the effectiveness of moderation decisions.

Procedural justice considerations also have implications for who should make moderation decisions. Community members often have mixed feelings about moderators' interventions. Members will be more positive about these authorities if they feel this power is deserved (through past contributions to the community or demonstrated expertise) or if the community had a say in the selection of these persons. Authorities who "deserve" their posts or who are selected by the community are more likely to be perceived as less biased and more likely to reflect the prevailing standards of the community than those who are self-appoint or appointed by site

owners. As they enact their roles impartially, they will be seen as more predictable and less biased.

For example, Slashdot.com instituted a moderation system (CmdrTaco, 2003), in which community members rate the quality of contributions to recognize poor or good contributions. The system was designed to "promote quality, discourage crap," that is, to encourage contributions that meet community's norms of quality. The moderation system is also community based to prevent a single moderator from exercising a "reign of terror."[1] Moderators cannot assign points to their own posts, to "prevent abuses." Finally, "to address the issue of unfair moderators," Slashdot created a 'meta-moderation' system, in which any logged in community member can evaluation the quality of others' moderation. Whatever other virtues this moderation system possesses, it increases community members' perceptions that the moderation process is fair and not subject to the capricious actions of just a few people.

Design Claim 4: Moderation decided by people who are members of the community, are impartial, and have limited or rotating power will be perceived as more legitimate and thus be more effective.

Production communities often employ quick reversion tools allow community members to repair damage done by vandals, newcomers or people who harm the product by mistake. Open-source software repositories use version control tools, such as Subversion or Git, for quick reversion, allowing administrators to easy rollback the code to a previous state when people offer buggy or inelegant code. Many wiki-based communities have tools to show differences between any two versions of a document and to instantly revert a document to a previous form. For example, the open-source, content management system Drupal and the open-source, wiki software MediaWiki, upon which many online communities are built, provides built in tools that allow users with a certain level of permission to revert any document (See Figure 1).

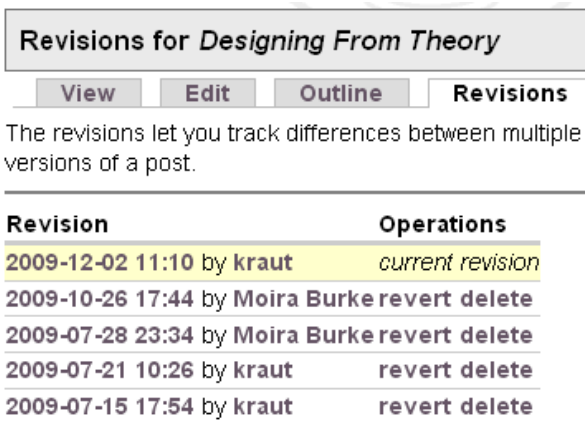


Figure 1. Reversion mechanism in Drupal

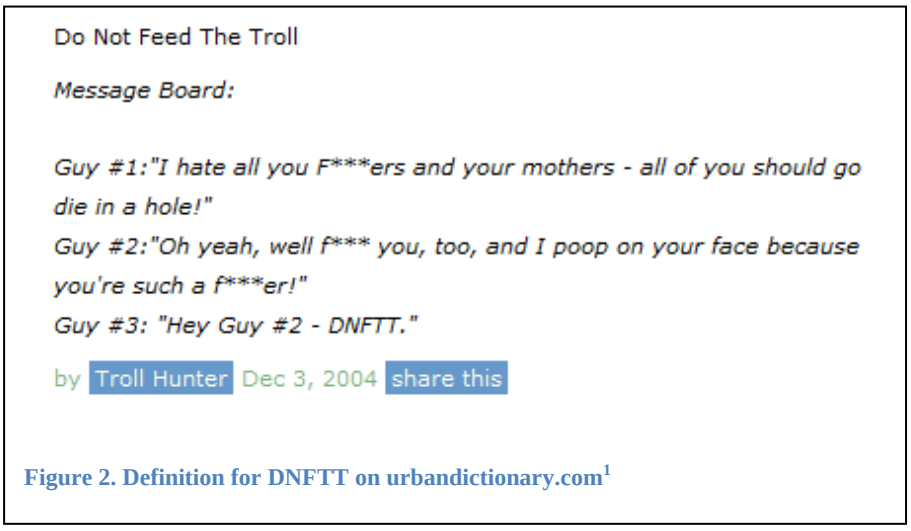
Design Claim 5: Reversion tools limit the damage disrupters can inflict in production communities.

In recommender systems, like Trip Advisor or MovieLens, where the threat is people trying to manipulate the recommendations that are made, the analog of moderation and reversion is to filter out or discount ratings suspected of coming from shill raters. Researchers have developed

algorithms that look for suspicious patters (e.g., too many in a short period of time, or insufficient variability). For a survey, see Mobasher et al (2007)(2007). Another approach, called the Influence Limiter, does not throw out suspect ratings completely, but partially discounts them: the discounting declines as the system gains confidence that the rater is honest rather than a shill for a manipulator(P Resnick & Sami, 2007). The problem with systems that partially discount or completely filter out ratings from suspect raters is that when they make mistakes, information from honest raters will not be fully utilized. An analytic model showed that such mistakes are inevitable: any system that limits the damage of shills will have to throw some information from honest raters as well (P Resnick & Sami, 2008)

Design Claim 6: Filters or influence limits can limit the damage of shill raters in recommender systems, but only at the cost of ignoring some useful information from honest raters.

One of the ways that trolls are able to disrupt is by eliciting reactions from community members that create strife within the community. For example, Herring et al describe how a troll in a feminist forum was able to provoke not only angry responses to him, but also disagreements among other members about whether his behavior was acceptable and what to do about it (Herring, Job-Sluder, Scheckler, & Barab, 2002). For a troll who is seeking to disrupt, sowing contention among other members is clearly a victory. Several group members argued that ignoring the troll would be more effective. However, doing so would have required everyone to recognize the troll and to follow a norm of ignoring him. As more people become experienced with participating in online communities, it may get easier for communities to follow a norm of ignoring trolls. Indeed, attempts to spread the norm have yielded an acronym DNFTT: Do Not Feed the Troll (see Figure 2).


Do Not Feed The Troll

Message Board:

*Guy #1:"I hate all you F***ers and your mothers - all of you should go die in a hole!"*
*Guy #2:"Oh yeah, well f*** you, too, and I poop on your face because you're such a f***er!"*
Guy #3: "Hey Guy #2 - DNFTT."

by Troll Hunter Dec 3, 2004 share this

Figure 2. Definition for DNFTT on urbandictionary.com[1]

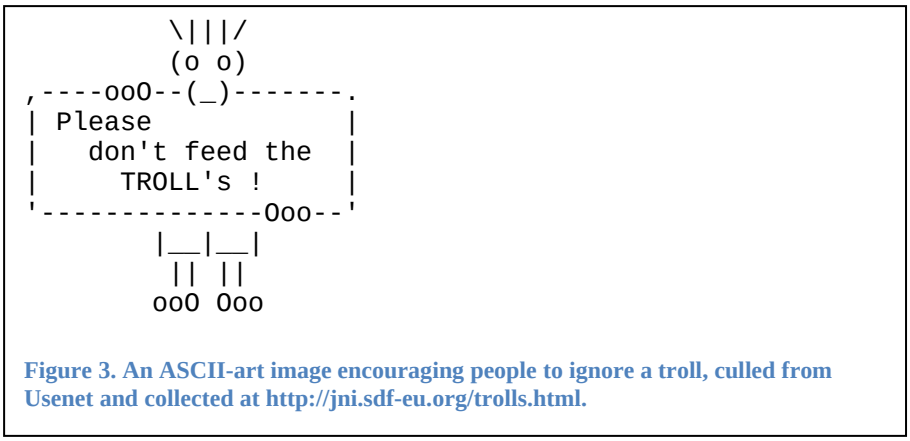


Figure 3. An ASCII-art image encouraging people to ignore a troll, culled from Usenet and collected at http://jni.sdf-eu.org/trolls.html.

Design Claim 7: A widely followed norm of ignoring trolls will limit the damage they can do.

Coerced Compliance: Limits on Bad Behavior

Even if each individual action can cause only limited damage, the damage can accumulate with lots of actions. Next, we consider ways to limit the amount of bad behavior that a bad actor can do.

Throttles or quota mechanisms are one way to prevent large-scale damage by a disrupter, especially damage caused by repetitive actions. For example, chat rooms can automatically block participants from posting too many messages in too short a time, or limit the number of links in those messages. The throttle prevents a person or program from barraging a community, whether intentionally, as in the case of a spammer, or unintentionally, in the case of an overzealous newcomer unaware of community norms. Similarly, when Facebook detects unusual speed or frequency of a behavior, such as friending other users or posting on their walls, it sends a warning to the user (Facebook, 2010). Though Facebook does not disclose its precise quotas, the warning message links users to relevant guidelines, such as how to promote a business or an event. Twitter lists spam-like activities that will lead to an account being investigated, removed from search, or terminated, including "aggressive follower churn," updates primarily comprised of links, or large numbers of duplicate @replies to other users (Crystal, 2009).

Design Claim 8: Activity quotas allow people to participate in a community, but prevent repetitive, spam-like activity.

Rather than responding to the quantity of activity, a member's activity may be limited based on a moderator's assessment of its quality. In PalTalk's chat rooms, for example, a room's current current owner can gag any of the participants. In other systems, a more severe gag or ban imposed on a member would apply community-wide, not just to a particular space. Gags and bans may be temporary, imposing a cooling period of a few minutes, hours, or days. Or they may be permanent.

As with moderating individual messages, gags and bans that are perceived as unfair may be resisted by the individual and his or her supporters in the community, and this can cause

significant disruption and damage as well. The considerations above about procedural justice and legitimacy of the moderators apply here as well, perhaps even more strongly. Gags and bans are often used as part of an escalating regime of sanctions intended to induce good behavior, a topic analyzed in more detail in the next section.

Beyond resistance that they may cause, gags and bans may be ineffective if they are easy to circumvent by using a different account. For example, in a chat room where people choose a handle to use as they enter the room, it may be trivial for someone who has been gagged to exit and come back a few seconds later. More generally, a gag or ban will stop misbehavior only as long as it takes to register for a new account.

When it is easy for people create new accounts, gags or bans may be more effective if the target is not immediately aware of the ban. There are several ways to disguise a gag or ban. For example, in a chat room, the gagged person may see an echo of everything she types, but her comments may not be displayed to others in the room. The gagged person may think that everyone is just ignoring her. Another possibility is to display a system error message suggesting that the site is temporarily out of service, but only show it to the gagged person.

Design Claim 9: Gags and bans can limit the continuing damage of a bad actor, but only if it is hard for the bad actor to use a different account or if the ban is disguised.

As with moderation decisions, a perception of procedural justice will make people think that gags and bans are more legitimate, and thus will more willingly acquiesce to them. For example, in Wikipedia, editors can be banned from particular pages or can be blocked from editing all pages. These actions are taken following a standardized procedure. Wikipedia's blocking policy lists types of behavior that warrants blocking an editor, evidence that someone needs to produce to request that an editor be blocked, specifies review by impartial administrators and an appeals process .

Design Claim 10: Consistently applied criteria for gags and bans, a chance to argue one's case, and appeal procedures increase the legitimacy and thus the effectiveness of gags and bans.

Another approach for limiting damage is to require people to earn the privilege of taking actions that are potentially harmful. Open-source software projects have explicit ladders of access. Although most projects allow anyone to post a bug-report to a public form, people who want to change code must go through a vetting process. Typically, they must send their patches or other small bug fixes to more senior developers, known as committers, before their code is integrated into the main software program. Only after they have shown a substantially history of offering high quality code and technical discussion are they granted committer status themselves (Ducheneaut, 2005; Krogh, Spaeth, Lakhani., & Hippel, 2003)

The Omidyar Network's community (Wikipedia, 2010b), where people discussed issues related to philanthropy, used an internal currency that could be spent to create new groups or discussions. The currency was acquired through participation in discussions, but was capped at three times the person's feedback score. This limited the ability to accumulate currency for those who participated in ways that others disapproved of.

The Influence Limiter for recommender systems, mentioned above, also instantiates this approach in the recommender system context [Resnick and Sami 2006]. New raters begin with a very small amount of reputation currency. Influencing others' ratings requires placing a bet: those without sufficient currency are limited in their ability to influence recommendations for others. Normal users who report on their actual opinions about items will naturally accumulate reputation currency, and thus influence on predictions, for other people who share their tastes. An attacker who has no information about the items being rated however, and employs some automated strategy for generating ratings, not revealing any real information about those movies, will, on average, not accumulate any currency with those fake ratings. The only way to gain currency, and thus influence, is to provide genuine information, which is easy for normal participants to do but hard for attackers to do.

Design Claim 11: Paying to take actions in the community with currency accumulated through normal participation will reduce the ability for trolls and manipulators to act.

Even if someone can do only a little damage with one account before being detected and stopped, if it is possible to new accounts automatically, the cumulative can be quite large. For example, the Influence Limiter described above gives a little bit of reputational currency, and thus influence to new raters, so that they can place bets and prove themselves. An attacker who can create thousands or millions of new accounts will be able to manipulate recommendations.

Figure 4: the reCAPTCHA service.

A CAPTCHA, which stands for "**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part" is a test presented to a computer user that should be easy for a human to pass but very difficult for a computer. Figure 4**Error! Reference source not found.**, for example, illustrates a CAPTCHA in which the distorted words from old scanned text, which are initially difficult to read and then rendered more difficult by adding additional distortions. By requiring applicants to complete a CAPTCHA before subscribing, the community can eliminate

automated spammers and other computer agents who are attempting to violate the community's norms[2]. Craigslist uses a related technique: it requires posters of classified ads to enter an email address and then respond to an invitation sent to that address before their ad goes public. Only bots that have access to a large number of distinct email addresses, using a variety of different email provider domain names, will be able to post large numbers of classified ads.

Rather than proving that they are human, account registrants may need to prove their identity, by providing a driver's license number or credit card number. Chapter 3, on newcomers, discusses methods for ensuring a good match between newcomers who are human and the communities they join.

Design Claim 12: Limiting fake accounts with CAPTCHAs or identity checks limits automated attacks.

Encouraging Voluntary Compliance

In addition to limiting damage and coercing compliance with behavior norms, people can be encouraged to comply voluntarily. Techniques for encouraging voluntary compliance tend to be more effective with insiders, who care about the community's health and their own standing within the community. To gain voluntary compliance with behavior norms, designers face two challenges. First, members of the community have to know the norms and be aware of them when making behavior choices. Second, members have to want to follow the norms, even when there are counter-vailing forces drawing them toward non-normative behavior.

Making Norms Clear and Salient

People learn the norms of a community in three ways:

1. observing other people and the consequences of their behavior;
2. seeing instructive generalizations or codes of conduct;
3. behaving and directly receiving feedback.

Psychologists distinguish between descriptive norms and injunctive norms. Cialdini defines descriptive norms as beliefs about typical behavior (Cialdini, 2003). Injunctive norms, on the other hand, define which behaviors people approve or disapprove of.

People tend to conform to descriptive norms, even though they lack the moral force of injunctive norms. The behaviors that others engage in may become a focal point, the first option that people consider. In addition, they may want to fit in by doing what others do. And they may interpret the descriptive norm, what people tend to do, as social proof of what the underlying injunctive

[2] Actually, CAPTCHAs do not eliminate the possibility of attackers creating many new accounts, but they do make it harder and a little more expensive. A New York Times article reports that sophisticated spammers are paying people in developing countries to answer captchas, with the going rate about $1 for 1000. The solved captchas are passed back to a computer program which automatically completes the rest of an account registration process (Bajaj, 2010).

norms are; indeed, in some circumstances, what people do may be a stronger indicator of what they truly believe is acceptable than any explicit statements they make.

In 1936, Muzafer Sherif put people in a dark room and showed them a pinpoint of light, which seemed to move anywhere from 1 to 10 inches. (This phenomenon is a perceptual illusion called the autokinetic effect.) After hearing other group members announce their estimates of how far the light moved, the group converged upon a norm, such as 3 inches, with individual group members' estimates varying in small amounts from this norm. Sherif's study was one of a long tradition of research into conformity—how people in groups learn what is acceptable behavior and adopt these norms without any external pressure. The power of observing others act in particular ways has been demonstrated repeatedly. The effect of the descriptive norms in the Sherif experiment can last over a year, even when individuals are tested individually, without the group being present. (Rohrer, Baron, Hoffman, & Swander, 1954).

Outside of experimental labs, as well, people's behavior produces signals about acceptable and unacceptable behavior to others. Erving Goffman's influential ethnographic studies of face to face interaction in mental hospitals, elevators, dinner parties, stadiums, and even casinos (where he became a skilled blackjack pit boss) described how people negotiate their way around often packed urban spaces, mark their territories while so doing, signal their relationships to others by various "tie-signs" and manage their appearances so as to appear normal or unremarkable(Goffman, 1959, 1963). Using a theatrical metaphor, he described how people act in ways that convey they can be trusted to act predictably within the range of acceptable behavior for their role. For instance, people eating alone in a restaurant often peruse a newspaper or paperback book to look occupied, because staring into space looks abnormal. When people fail to look normal, their behavior signals moral failure and can lead to their being stigmatized.

It follows from this discussion that one way to encourage normative behavior is to make others' normative behavior visible to all members of the community. Highlighting descriptive norms can change behavior. Many colleges and universities use social norm marketing to attempt to reduce heavy and binge drinking among their students by publishing accurate information about how much the typical student drinks. In these campaigns, the universities conduct surveys to identify the actual amount of drinking on campus and then advertised these rates via posters, direct mail, campus news paper and other means. Studies have shown that the actual rates are lower than many students think (Perkins, Meilman, Leichliter, Cashin, & Presley, 1999). DeJong and his colleagues conducted a large randomized field experiment of the effectiveness of these social marketing programs at 18 universities. Compared to universities that were not assigned to conduct a social norm marketing campaign, students in universities randomly assigned to participate in campaigns increased the the accuracy of their estimates of the others' drinking behavior on campus and decreased their own drinking, and these effects were stronger the more intensely the university participated in the campaign (DeJong, et al., 2006).

Online community designers have a number of options for highlighting the typical behavior in the group (i.e., the descriptive norms). At one extreme, they can simply make samples of individuals' actual behavior in the community visible to others. If the behavior is relatively homogeneous and the typical behavior is also the desirable behavior this technique should lead others to act similarly, in desirable ways. Buyers and sellers on the online auction community eBay can leave each other positive and negative comments. Both buyers and sellers overwhelming give positive feedback (about 99% positive approval ratings according to C Dellarocas & Wood, 2006; P Resnick & Zeckhauser, 2002). Thus, when members of the community browse the site, they are likely to see that others are typically responsible, and they will be motivated to be responsible themselves. Although the comments typically posted on news feeds on Facebook are more varied than the feedback sellers receive on eBay, most are casual and benign (e.g., "…has a cold and skinned knees. Totally reliving third grade."). Therefore, these examples of public user behavior provide descriptive norms for the prevalent behavior and should encourage similar casual and benign conversation.

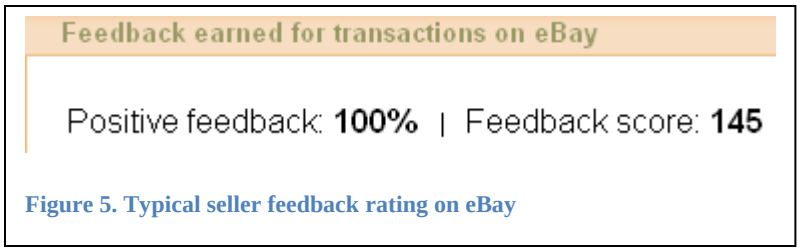


Figure 5. Typical seller feedback rating on eBay

Rather than depending on random encounters, designers can also choose to highlight particular exemplars of desirable behavior. For example, in a forum-based community, a "post of the week" could be highlighted on the front page.

Design Claim 13: Publicly displaying examples of appropriate behavior on the site will show members what is expected and increase their adherence to those expectations.

Showing members a small sample of norm violations can encourage appropriate, normative behavior, if it contrasts with the clearly more prevalent descriptive norm. According to Cialdini's focus theory, people learn norms from salient behaviors, actions that stand out and point people to what is appropriate to do in a situation. Abstractions and routine behavior can be hard to make salient but negative behavior catches people's attention. Negative examples thus may highlight the background norm. Cialdini

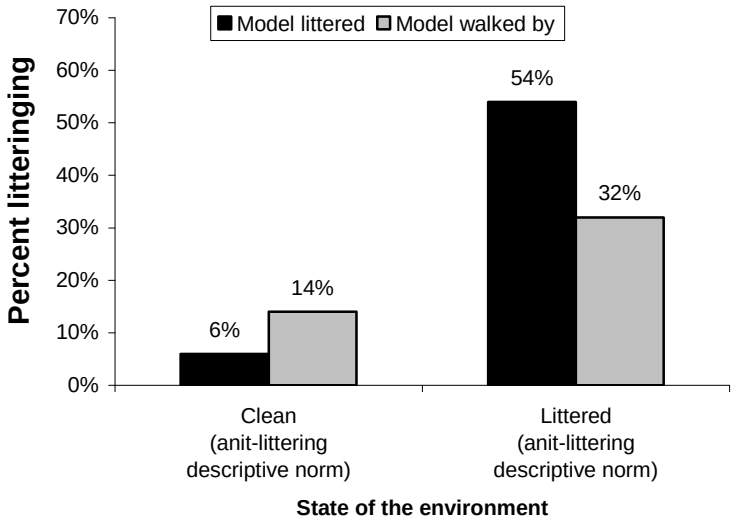


Figure 6. The effect of behavioral exemplars on highlighting discriptive norms

and his colleagues have shown that when people see a model litter in the context of an already littered environment they litter more than if they hadn't seen the model; However, seeing the same littering behavior in the context of a clean environment causes people to litter less than they would otherwise (SeeFigure 6.) (Cialdini, Kallgren, & Reno, 1991, study 1). That is, a negative example that violates a descriptive norm makes the norm more salient and causes more people to act consistently with it.

In the context of online communities, one example of non-normative behavior may bring into clearer focus a pattern of normative behavior. Many online communities such as discussion boards or wikis preserve records of misbehavior and make them salient. For example, when moderators flag or disemvowel a message, or respond to it with a suggestion that the conversations move elsewhere, there is still a visible trace that non-normative behavior occurred. When a message is moved or removed without leaving a trace, others will no longer know that a violation occurred.

Wookieepedia, a wiki dedicated to Star Wars (Wookiepedia, 2010), has a prominent page on being civil and respectful, which includes examples of personal attacks that should be avoided. Sometimes conflict continues to escalate and mediators are brought in to help resolve it, in which case records of such conflicts are preserved for the community to potentially reference in the future.

Design Claim 14: Publicly contrasting examples of inappropriate behavior in the context of a descriptive norm of appropriate behavior will highlight the descriptive norm and increase people's adherence to it.

This tactic of highlighting non-normative behavior can backfire if it leads to the impression that the behavior is engaged in by large numbers of people and thus is in fact a descriptive norm. In the experiment described above, seeing someone litter in an already littered environment led even more subjects to litter than those who simply observed the littered environment. As another example, the attempt by the National Park Service to deter theft of petrified wood using a sign that reads "Your heritage is being vandalized every day by theft losses of petrified wood of 14 tons a year, mostly a small piece at a time." may in fact be causing visitors to steal more and his colleagues conducted an experiment comparing two signs designed to deter this theft (Cialdini, 2003). Both signs urged people not to steal. One, though showed three thieves and had the message "Many past visitors have removed petrified wood from the Park, changing the natural state of the Petrified Forest", while the other showed a single thief and simply said "Please don't remove the petrified wood from the Park, in order to preserve the natural state of the Petrified Forest." Visitors were four times more likely to steal after seeing the sign with three thieves than the sign with one.

These results suggest that in a community like eBay, where the descriptive norm is one of honesty, revealing a dishonest seller should increase members' honesty not undermine it. However, highlighting bad sellers could backfire if the background prevalence of such sellers was higher. Similarly, if there are many people in a hobbyist community who, against the stated rules of the community, post advertisements for their businesses, highlighting these norm violations may only serve to embolden other members to advertise their businesses as well.

Thus, community managers face a difficult decision, when they take administrative action to remove content or move it to a more appropriate place, of what traces to leave behind about those decisions. On the one hand, seeing a trace, and being able to follow a link to find details about what was moved or removed, and why, may help people to learn the administrator's norms of behavior. On the other hand, seeing many such traces will suggest that the administrators' expectations are not a consensus of the community. The best decision will often depend on the prevalence of the non-normative behavior; perhaps counter-intuitively, the more of it there is, the less trace of its existence should be left visible.

Design Claim 15: Publicly displaying many examples of inappropriate behavior on the site will lead members to believe this is common and expected.

While observing common online behavior illustrates the descriptive norm (i.e., behaviors that are common), observing responses to those behaviors teaches the injunctive norms (i.e., what behaviors are approved or disapproved). Observers need to see the consequences of behavior, for instance, the feedback that others provide to it, to understand its appropriateness. Do others say thank you for behavior x and not for behavior y? Do they rate person x higher than person y? Does one person seem to have a better reputation than others? Seeing behavior along with its consequences makes norms more learnable (E. Fehr, Fischbacher, & Gächter, 2003).

The feedback can be informal, formal or both. The informal feedback that one member of a health support group provided others who answered her questions clearly highlights valued behavior: "*I want to thank all of you who responded to my posting. I dont know what my future will be with regards to diabetes, but knowing that there are people out there who care about each other is wonderful. I especially want to thank …"*[3] eBay's feedback mechanism combines informal comments (e.g., "Great honest ebayer! I would purchase from again tomorrow!") with formal, symbolic feedback (positive, neutral or negative ratings).

Research suggests that formal feedback is more effective that informal feedback in helping people learn the norms of appropriate behavior. Moon and Sproull (2008) compared technical support groups for software problems that allowed only ad hoc member feedback about the quality of contributions in the text of replies to support groups that allowed more formal feedback (e.g., awarding points or stars). Consistent with a reinforcement model, where quality answerers who get systematic feedback contribute more and lower quality answerers who get systematic feedback improve or drop out, they found that the formal approach was more effective than informal feedback: technical problem resolution was more effective, and that people who had higher quality contributions had longer participation duration.

Design Claim 16: Displaying feedback of members to others increases members' knowledge of community norms and compliance with them; formal feedback is more effective than informal feedback.

[3] Quoted from alt.support.diabetes. We don't provide a citation for this quote to protect the poster's privacy

Inferring a norm from a sample of behavior in a community can be difficult when there are many examples to observe and they vary in the extent to which people adhere to the norm. In large and active communities, there may simply be too much to look at to get a sense of what is appropriate in the community simply by looking at samples of behavior. To convey a descriptive norm, one alternative is to display easily interpreted statistics tallying certain types of behavior. For example, just as some workplaces prominently display a sign showing the number of days since the last workplace injury, a community could display the number of messages since the last reported abuse, or the (low) percentage of messages flagged for violating the community's official policies.

Design Claim 17: In large communities, displaying statistics that highlight the prevalence of normative behavior will increase members' adherence to normative behavior.

**** ETIQUETTE ****

A note about etiquette. Keep in mind when responding to a topic or entering a new one that the other users also have feelings. Please avoid trampling on them. Also, remember that comments entered in hasty reaction to someone else's posting will be available to be read long after you have entered them. So it is wise to exercise some moderation and good judgment.

Figure 7. Early behavioral guidelines from the Well, 1985

When observers need to infer the norms by integrating over many, varying examples of behavior, it is often helpful to crystallize the generalization process by providing community members with explicitly stated guidelines or rules. These statements may be descriptive (e.g., "Generally, we are nice to each other even as we critique each other's photos") or injunctive (e.g., "Be nice even as you critique someone's photo." They may either describe normative behavior (dos) or non-normative behavior (don'ts). They may be set at the origination of the community, or articulated as a response to critical events in a community's history. Often, a rule is made in order to settle an argument about whether someone's behavior violates a norm or not. For example, Lambda Moo's norms about violence against others' avatars were codified only after the cyber-rape incident described earlier.

The Well's Host Manual was an early attempt to create community guidelines for The Well, one of the first online communities founded in 1985 (Hoag, 1996; Williams, 1997). Community designer Amy Jo Kim has a webpage that provides sample codes of conduct and rules for community businesses and associations at www.naima.com/community/policies.html. Another list is at www.fullcirc.com/community/sampleguidelines.htm.

Research evidence indicates that when norms are clearly stated, people are more likely to act consistently with the norm, and do so over a wide variety of situations. Norms that can reasonably be inferred but are not directly stated may not be noticed, understood or obeyed. For example, motorists returning to their cars were less than likely to toss on the ground a handbill stuffed under their windshield wiper when the handbill explicitly reminded them not to litter ("April is Keep Arizona Beautiful Month. Please Do Not Litter.") than when it urged a related action ("April is Preserve Arizona's Natural Resources Month. Please Recycle."), as shown in Figure 8 (Cialdini, et al., 1991, study 5).

The effects of making a social norm explicit are stronger when the norm itself is less clear. For example, Zitek and Hebl found that participants were far less likely to condone prejudice against others when they heard another person also condemn it and more likely to condone it when they heard others condone it compared to conditions when they heard nothing (Zitek & Hebl, 2007). These effects of hearing the norm made explicit were stronger when pre-existing social norm was more ambiguous (discrimination against ex-convicts and racists) compared to groups where the pre-existing norms were clear cut (discrimination again blacks and gays).

Norms are often less clear in the early stages of a new community or any time there is fast growth. When reflecting on the current state of Wikipedia and lessons he learned as one of its founders, Larry Sanger expressed regret that the current Wikipedia community does not sufficiently defer to experts and specialists when they write in their areas of expertise, even though the community did so during the encyclopedia's early days.

> *"This is just common sense," as I wrote, "but sometimes common sense needs to be spelled out." What I now think is that that point of common sense needed to be spelled out quite a bit sooner and more forcefully, because in the long run, it was* not *adopted as official policy, as it could have been. -- (Sanger, 2005, p. 318)*

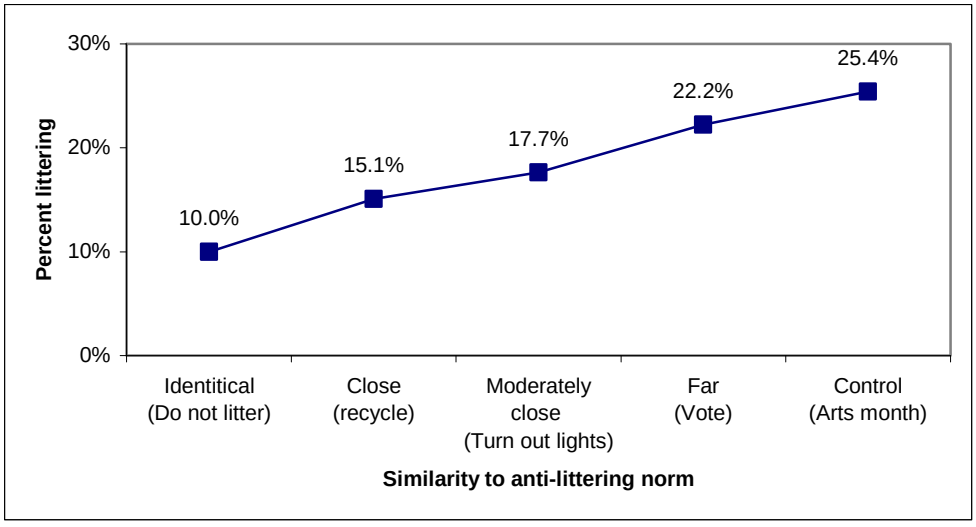


Figure 8. Effects of norm explicitness on compliance (adapted from Cialdini et al, 1900)

Design Claim 18: Explicit rules and guidelines increase the ability for community members to know the norms, especially when it is less clear what others think is acceptable.

Another important design decision is how prominently to display guidelines and rules. Some communities require newcomers to read particular rules before they join or post, or post them prominently where everyone will see them frequently. Others do not.

Social news aggregation sites such as Reddit (www.reddit.com) face a special challenge in making norms and rules salient since the entire content of these sites revolves around voting and commenting on web links. Although Reddit has an area where rules are articulated (known collectively as "reddiquette"; http://reddit.com/help/reddiquette?), this area has low salience for most users.

The Reddit community's solution is to post dummy "articles" in the main news area whose titles describe the norm or rule. Those "articles" that have widespread support and relevance in the community are voted up, often reaching the front page and thus becoming highly salient. An example of this is a post advocating using comments for conducting polls instead of articles:

> From reddiquette: "Please don't conduct polls using posts. If you feel you must use Reddit to conduct a poll do it using a comment. Create a self referencing post and then add a comment for readers to mod up or down based on their answer to your poll question. Also, be sure to indicate in the title of your post that the poll is being conducted using comments. Including something like "(use comments to vote)" in the title would probably be sufficient."

This new rule was developed in response to a slew of polls taking over the front page of Reddit, as each poll "vote" had the side effect of increasing the poll's popularity and visibility. The new "article" garnered widespread support and high salience (it was voted up more than one thousand times), at one point reaching the #1 article spot.

Unfortunately, prominently displayed rules and guidelines can convey a negative impression. Potential members may fear that they will not be able to do what they want to, or that they will accidentally fall afoul of one of the rules that they didn't understand. Newcomers to Wikipedia may inadvertently violate one of its many policies and guidelines designed to regulate behavior in "the encyclopedia that anyone can edit'. As a result, their contributions are frequently reverted, and they become discouraged or driven away (Halfaker, Kittur, Kraut, & Riedl, 2009, Figure 5). When we have assigned students to edit a page of their choice on Wikipedia, and tell them to read Wikipedia's guidelines first, many report feeling very intimidated about making an edit.

Paradoxically, prominently displayed or excessively detailed rules may also convey the wrong descriptive norm. A natural inference is that the rules were created in response to problematic behavior. It is also natural to infer that such behavior must occur fairly frequently, else it would not be necessary to prominently display the rules.

Design Claim 19: Prominently displayed guidelines may convey a descriptive norm that the guidelines are not always followed.

Many sites have compromised by creating explicit rules and guidelines, but burying them deep in the site, visible only to people who go looking for them. Gaia.com is a community site for teens 13 and up. It is a very challenging regulatory environment due to the concern of many parents and many laws to protect children. The Gaia site has developed an extensive list of rules, safety tips, and information for parents. Yet teens have a strong dislike of rules and probably would run in the other direction if the rules were thrown in their face. So, the rules are there if you look for them, but they are not prominent. While most members are probably unaware of these explicit rules, as we shall see below, even rules that are not noticed until people are pointed to them may have some value in creating legitimacy for assertions of norms or sanctions for violating them.

Another option is to make rules and guidelines prominent, but only at the point where people may be about to violate them. For example, eBay has a guideline that buyers should try to resolve conflicts with sellers before leaving negative feedback for them. This guideline is brought to the attention of buyers when they are about to leave negative feedback.

In a community where there was a strong norm of polite, supportive responses, automatic text analysis of posts that are submitted could be used to alert people that they might be about to violate that norm without forcing everyone to read about the guidelines. Similar features have been built into some email programs. For example, Eudora's MoodWatch software, invented by David Kaufer, automatically cautions emailers who are about to send a message containing "flame" words (Shankland, 2000, September 15).

Design Claim 20: Offering people reminders at the point of an action that may violate norms will reduce the number of offenses.

Enhancing Compliance

To this point, we have argued that people must know the norms of a community before they can follow them and have suggested a number of design choices that should increase this knowledge. Even when they are aware of the norms, however, people may not always comply. Four things will increase compliance: commitment to the community, legitimacy of the norms, the ability to save face, and expectations about rewards for compliance or sanctions for non-compliance.

Scholars since at least the time of Durkheim have argued that group cohesion contributes to the social order (Durkheim, 1953 [1903]). To the extent that community members care about the welfare of the community and see the norms as linked to that welfare, then community members are more likely to comply with norms the more they identity with the community and to enforce the norms. Chapter 4 on encouraging commitment to online groups examined ways of promoting cohesion. As described in more detail in that chapter, designers can promoted cohesion by emphasizing group identity though providing distinctive group names and missions, emphasizing group interdependence and competing against out-groups. Alternatively, one can increase group commitment and cohesion by emphasizing the interpersonal relationships between individual group members, for example, by keeping group sizes small, by creating opportunities for members to repeatedly see and find out about each other, and by encouraging interpersonal communication and mutual self-disclosure.

Design Claim 21: In more cohesive groups to which members are more committed, members will be more likely to spontaneously comply with the norms.

Externally imposed rules and monitoring tend to be viewed as unfair and to lead to conflict. From Ostrom's studies of successful institutions for governing common pool resources, design principle #3 was "Collective-choice arrangements. Most individuals affected by the operational rules can participate in modifying the operational rules." Ostrom (2000) argues that collective choice leads to rules that are better tailored to specific situations, but also that it builds legitimacy and thus compliance with the rules. Even if the group spends more time initially in discussion and comes to the same decision in the end as that made by an elite core, involving them in the decision making process should result in long-term benefits.

In early May 2007, a member of the news aggregation site Digg.com posted a news story consisting of a HD-DVD cracking key – a set of hexadecimal numbers that provided people a way to circumvent DVD copyright protection. In a rare display of censorship, Digg.com's administrators removed the post. However, instead of accepting this enforcement of rules from above, community members reposted the key over and over until the Digg homepage was little else than posts about the key and stories about how the Digg leadership had "betrayed" the members of the site. The disruption to the site was accompanied by an exodus of disgruntled members, with Reddit – a competitor to Digg – including a top story welcoming the Digg exiles.

Everquest's management imposed rules when it perceived bad behavior was driving away subscribers. (Multiple player gaming environments have experienced many hacker attacks, destruction of property, flame wars, and spirals of retaliation and cross-retaliation, see Kolbert, 2001. In Yee's 2001 survey, 20% of the respondents answered "yes" to the question, "Would you hack the game if you could?") Rules were necessary, but only 12.5% of Everquest members thought that management's top-down "Play-Nice" rules helped the environment. Social norms and rules generated by the community might have worked as well as external rules and would have more staying power.

On the other hand, LambdaMOO, one of the first true online communities, used community rule making, with good results:

> *". . . we started having disagreements about what was and was not proper conduct here. Eventually, I was approached by a number of players and asked to draft a set of rules for proper MOO behavior. . . I showed the draft to a bunch of people and asked for their comments on its style, completeness, and correspondence with their impressions of the 'right' way of things. After incorporating suggested changes, the first version of 'help manners' was publicized in the newspaper; I had, I think, done as good a job as I could of trying to capture the public consensus of that (admittedly early) time. Perhaps surprisingly, 'help manners' worked quite well in reducing the number of incidents of people annoying each other. That society had a charter that reflected the general opinion and social pressure worked to keep the MOO society growing fairly smoothly." [from a 1996 post, "LambdaMOO Takes a New Direction" in the LamdaMOO help system]*

Design Claim 22: Community influence on rule making will increase compliance with the rules.

When people violate community norms and the violation is brought to their attention, they will be much more willing to discontinue the bad behavior and correct previous errors if they can do so without having to admit that they deliberately violated the community's norms. If they can plausibly claim ignorance, or that their actions were misunderstood, or that the action was not theirs, they can save face.

For example, in the mid-1990s, MIT adopted procedures that they called "stopit" for dealing with harassment that occurred through computers on campus. A key element was a procedure for notifying norm violators that gave them a face-saving way out. As Gregory Jackson, then Director of Academic Computing, wrote,

> *The third stopit mechanism is a carefully-structured standard note to alleged perpetrators of harassment, improper use, or other uncivil behavior. "Someone using your account," the note begins, "did [whatever the offense is]." The u.y.a. note (as this mechanism is known, for its introductory words) then explains why this behavior or action is offensive, or violates MIT harassment policy, or Rules of Use, or whatever. "Account holders are responsible for the use of their accounts. If you were unaware that your account was being used in this way," the note continues, "it may have been compromised. User Accounts can help you change your password and re-secure your account." Detailed directions to User Accounts follow. The note concludes with a short sentence: "If you were aware that your account was being used to [whatever it was], then please make sure that this does not happen again."*
>
> *Two interesting outcomes ensue. First, many recipients of u.y.a. notes go to User Accounts, say their accounts have been compromised, and change their passwords - even when we know, from eyewitnesses or other evidence, that they personally were the offenders. Second, and most important, u.y.a. recipients virtually never repeat the offending behavior. This is important: even though recipients concede no guilt, and receive no punishment, they stop. If we had to choose one lesson from our experience with misbehavior on the MIT network, it is how effective and efficient u.y.a. letters are. They have drastically reduced the number of confrontational debates between us and perpetrators, while at the same time reducing the recurrence of misbehavior. When we accuse perpetrators directly, they often assert that their misbehavior was within their rights (which may well be true). They then repeat the misbehavior to make their point and challenge our authority. When we let them save face by pretending (if only to themselves) that they did not do what they did, they tend to become more responsible citizens with their pride intact (Jackson, 1994).*

There are other face-saving mechanisms, besides the "someone using your account" phrasing. Another possible phrasing for a notification message is something like, "You may not have been aware of this guideline, but we have a stated policy of [fill in here]. Please see [link to policy]. No big deal, but please stick to this in the future." Rather than allowing people to claim that it was someone else, it allows them to plead ignorance.

Giving people the option of undoing their offending action without leaving a trace of it also helps people to save face. For example, suppose someone makes a post that violates a community norm, and someone else posts a response chiding the original author. If the original author has the opportunity to remove both the original post and all replies to it, leaving no ongoing source

of embarrassment, he may do so willingly. Without that option, he may feel the obligation to defend the action, and even repeat it to demonstrate that he thinks it acceptable.

Design Claim 23. Face-saving ways to correct norm-violations will increase compliance.

Rewards and Sanctions

Both classic and contemporary theories of deterrence in criminology can help designers think through the best way of preventing misbehavior (Gibbs, 1985; Pratt, Cullen, Blevins, Daigle, & Madensen, 2006a). These theories hold that the decision to commit a crime or more generally to violate a norm is in part a rational decision. Although people vary in their predisposition to commit crimes (e.g., by class, geographic area and race), deterrence theory argues that those with criminal disposition will violate the rules only when it "pays." That is, based on an informal cost benefit analysis, they perceive that benefits outweigh the costs. Thus, the argument is that actual punishment for offenders and the threat of punishment for future offenders deters misbehavior. Researchers have studied how various factors associated with the threat of punishment, such as the use of warnings, and the certainty of punishment, its swiftness, or its severity, deter violations of norms and rules.

Sanctions may be delivered by community members but outside the online community. For example, people who send inappropriate messages to a school-wide email list may be shunned when they encounter fellow students in classrooms or hallways. More typically, external sanctions are not possible and sanctions are delivered within the community. For example, members may be publicly scolded, or their posts may get low ratings. Their contributions may be reverted, or their messages deleted or moved. They may lose privileges such as posting in particular areas or committing code in open source projects. They may be shunned within the community: in a gaming community that detracts from the fun, in a commerce community like eBay it subtracts from the profits. They may even be banned from particular activities or the community as a whole, temporarily or permanently.

For spammers, manipulators, and trolls, simply limiting the effectiveness of their actions reduces the incentive to participate. Thus, all the techniques described as ways to limit the damage they do, if effective, will also help to reduce their incentives to try. For example, following a norm of not feeding the trolls not only limits the collateral damage their behavior can cause, but also makes it leass appealing for them to participate in the community.

For spammers, reducing the chances that their posts will be seen has a similar effect. In the particular case of link spam, bots post links in online forums and blog comments, to commercial (often porn) sites. The spammers' real audience is not the readers of the forums or blogs but search engines that crawl the forums and blogs specifically looking for links. Most search engines give higher rankings to sites that have more incoming links (see, for example, the PageRank algorithm that was the initial inspiration for Google's search engine (Page, Brin, Motwani, & Winograd, 1998). Many blog platforms, including Blogger and WordPress, while allowing newcomers to post comments subject to the blog owner's preferences, they also automatically include the "rel=nofollow" attribute in links embedded in comments. This mechanism directs search engines not to trust these links, preventing spam links from receiving PageRank, and thus discouraging spammers from disguising links to their products within blog

comments. Slashdot also uses the nofollow attribute in comments from potentially misbehaving users', using heuristics based on the age of the user's account and the user's karma (Wikipedia, 2010a).

Design Claim 24: Telling search engines not to follow links will discourage spammers from posting links.

For people, the simplest form of sanction is social approbation from other people. People are very sensitive to the public impression they give off to others (Goffman, 1959). Their concern about 'looking good' and how others will evaluate them often causes them to under-report lying, stealing drug use and illicit sexual relations in in-person interviews compared to anonymous surveys (S. Kiesler, Weisband, & Drasgow, 1999), to comply with experimenters' expectations in psychology experiments (Rosenberg, 2009), to give more money to charities when the identities of contributors are revealed (Alpizar, Carlsson, & Johansson-Stenman, 2008; Soetevent, 2005), and work harder in group setting when others know the identities of the contributors (Karau & Williams, 1993). These effects all depend upon people believing that other can see their behavior and identity them with it.

Festinger and his colleagues defined *de-individuation in a group* as being submerged in it. The individuals zre not seen or paid attention to as individuals and do not feel that they stand out as individuals (Festinger, Pepitone, & Newcomb, 1952). Zimbardo's experiments on deindividuation suggest that behavior that people are more likely to violate established norms when they can conceal their identities under white robes. A systematic review of 60 separate experiments indicates that de-individuation encourages anti-normative behavior, although the effect is not a strong one (Postmes & Spears, 1998). Anonymity, at least to outsiders, and larger group sizes both lead to more anti-normative behavior.

For the reasons we have just described, identifiable individuals may be more likely to adhere to group norms than anonymous individuals, especially when they face social sanctions for misbehavior (Sassenberg & Postmes 2002). Research suggests that the relative anonymity of online communication compared to phone and face-to-face communication is partially responsible for reduced normative pressure online (Bordia, 1997) as revealed, for example, by more flaming and other incivilities online (S. Kiesler, Zubrow, D., Moses, A., & Geller, V., 1985). For example, online, people are more willing to lie about themselves to potential romantic partners (Cornwell & Lundgren, 2001). The observation that 97% of vandalism to Wikipedia articles is done by anonymous editors is also consistent with this rationale (Wikipedia, 2010c).

Therefore one way to increase people's willingness to comply with the norms of a community is to prevent anonymous participation. For example, Wikipedia requires editors to register before they can edit some especially contentious pages. Sites where misrepresentation is a problem often require verified or third party authentication of identity for anyone who could potentially harm others. Many dating sites let anyone peruse the site but require driver license photos in exchange for the email addresses of other members. In 2009 Twitter began verifying the identities of well-known users, giving them a badge on their pages that serves to confirm that

they are who they say they are. These communities use authentication of identities to discourage potential harm doing by community members.

Design Claim 25: Verified identities and pictures will reduce the incidence of norm violations.

Identifiability—the ability for others to see and judge actions and associate them with the actor—encourages good behavior and discourages bad behavior in the moment. That is, people are concerned about how others will judge them even in single-shot encounters, where they do not expect to interact with the same people in the future. But concern about future interactions can enhance the power of these social judgments. For example, in laboratory experiment, people conform more to group opinions when they anticipate future interaction with members of the group (Lewis, Langan, & Hollander, 1972). A person's actions affect her reputation, and thus how people will interact with her later: 'the shadow of the future', as Axelord calls it (1985, p. 232), creates an incentive for good behavior in the present.

Informal tracking of reputations sometimes yields only a small shadow, however, for two reasons. First, some actions are not publicly observable. For example, on eBay, a seller may misrepresent the goods she is selling. The buyer will recognize it and refuse to buy from the seller again, but the seller's ability to sell to other buyers will be unhindered unless the unhappy buyer has a way to communicate with other potential future buyers. Second, there may be so many actions that it is hard for people to judge someone's overall reputation.

In online communities, explicit reputation systems can help to solve these problems (P Resnick, Kuwabara, Zeckhauser, & Friedman, 2000). For example, eBay provides an opportunity for buyers and sellers to leave comments about each other after their transactions. These comments are visible to others in the future, providing a public window onto the previously private transaction. Second, eBay provides summary statistics so that people do not have to read all of the individual comments and ratings. A potential buyer can quickly read that one seller has 99.5% positive feedbacks while another has only 94%, a difference that would take much longer to assess by looking at pages of individual feedbacks.

Empirical evidence suggests that explicit reputations can be an effective sanctioning mechanism. For example, in cross-sectional comparisons of naturally occurring transactions on eBay, more positive feedbacks led to higher prices and probability of sale, and the opposite for negative feedbacks(Bajari & Hortacsu, 2004; Cabral & Hortacsu; Dewally & Ederington, 2006). In other communities, the deferred reward or sanction created by reputations may not be economic in nature but instead affect future interactions in other ways. For example, in the Omidyar.net community described earlier, reputation scores had an impact on members' ability to create new workspaces and discussion threads. Perhaps more importantly for members who were psychologically invested in the community, each person's reputation score was displayed next to each post they made, and high scores became a valued status marker. [4]

[4] The Omidyar Network is a charitable foundation set up by eBay founder Pierre Omidyar. It is probably not a coincidence that the Omidyar.net community thoroughly integrated a reputation system into its operations.

Design Claim 26: Reputation systems, which summarize the history of someone's online behavior, help to encourage good behavior and deter norm violations.

Rather than depending on reputational consequences to affect someone's future interactions in the community, rewards and sanctions can be charged directly to someone's account. Rarely, money might change hands. More commonly, accounts might be denominated in some internal currency that is earned through actions taken within the community, such as the one used in the Omidyar.net community. If prices are higher for undesirable actions than for desirable actions, people will do fewer undesirable actions. The problem is assigning prices when actions of the same form (e.g., posting a message) may be desirable or undesirable depending on their contents. One solution is to assign the prices after the fact, based on feedback from other users.[5]

For example, Van Alstyne proposes that email senders should post a small "attention bond" (Van Alstyne, 2007). Recipients who were unhappy to have received a message would have the right to collect that small fee. Those who thought the message was reasonable to send would return the fee. People sending direct messages to individuals would face little risk but senders of commercial spam to thousands or millions of people might end up paying quite a bit to do so.

The Influence Limiter for recommender systems, described previously, works analogously (P Resnick & Sami, 2007). Each rating that changes the predictions that are made for other people is treated as a bet that other people will or won't like various items. When the other people enter their ratings for those items, the bets are resolved and people either gain or lose currency. Bad ratings, those that move predictions in the wrong direction, are costly, while good ratings actually earn currency. The amount of influence a person's rating has on a prediction determines the amount of the bet made, and bet sizes are limited by the currency holdings of the bettor. People whose ratings have been very helpful in the past get more influence on predictions for other people.

Design Claim 27: Prices, bonds, or bets that make undesirable actions more costly than desirable actions will reduce misbehavior.

While identifiability creates opportunities for both informal and formal reputational sanctions, many communities allow either completely anonymous participation, or participation under a long-lasting pseudonym that is not linked to an identity outside the community. For communities where members might not want to reveal their participation publicly, such as an HIV/AIDS discussion group or an activist political group fearing government repression, it is clear why members would prefer anonymous or pseudonymous participation. But even in many other, somewhat less sensitive arenas, people often prefer not to reveal their true identities, in order to

[5] If we think of reputation scores as account values in some internal currency unit, then these currency charges are actually a form of reputation consequence. The difference is that currency units other than reputations may have consequences other than changing how other members perceive the currency holder.

preserve some separation of context between different aspects of their lives. For example, one of the authors goes by the name "informationist" on eBay.[6]

Pseudonyms are popular in online communities. However, the impact of any kind of sanctions, including reputational consequences, for bad behavior is muted if someone can simply create a new account and start over. This is often referred to as the problem of cheap pseudonyms.

Cheap pseudonyms are especially problematic for sanctions such as bans: as mentioned previously when bans were discussed as a means of limiting bad behavior, people who are banned from an online community can come back with a new account name and thus escape the consequences of their previous actions. If the community wants to be open to newcomers, the norm violator who returns under a new pseudonym will have the opportunity to violate norms again before being banned again. And again. And again. As Donath pointed out, eventually the community may become less open to newcomers, not giving them a chance to violate norms until they have proved themselves, but that may be a bad outcome for the community as well, as valuable newcomers may be turned away by the need to prove themselves (Donath, 1997) (Donath 1999, p. 54).

The same problem can occur even if the sanction is not a complete ban. For example, consider a reputational consequence on eBay. Suppose buyers were very severe in their interpretations of seller feedback profiles. Suppose buyers were willing to spend $50 for an item from a seller with an unblemished record of only positive feedback from 100 or more transactions, but only $40 to buy the same item from a seller with one negative feedback, and $45 to buy from a new seller. This would not be sustainable. After receiving a negative feedback, a seller would choose to start over, and make $45 on her transactions rather than accepting the $40 she would receive from continuing to sell with an unfavorable feedback profile. If starting over is cost-free, then the worst possible seller feedback profile that anyone will continue to use will have to be treated by buyers no worse than they treat new sellers.

Friedman and Resnick created an analytic model that helps to clarify the predicament created by cheap pseudonyms and some of the strategies for dealing with it. The fundamental constraint is that for someone who is sanctioned, the utility of continuing to participate under the current identity, accepting the sanction, must be higher than the utility of starting over with a new pseudonym. Even moreso, then, the utility of participating with an established identity without sanctions must be higher than the utility of participating as a newcomer.

This formulation suggests three ways to maintain effective sanctions given the possibility of cheap pseudonyms. The first is to increase the benefits of maintaining a long-term pseudonym. Online role-playing games such as Everquest and World of Warcraft naturally include such a mechanism, as online characters need to be leveled up in order to gain access to new realms, equipment, or capabilities. When creating a new character, players lose these assets and can no

---

[6] Of course, revealing this partially collapses the context separation between my professor persona and my trader persona. Interestingly, the collapse is largely one way. Readers of this book can look me up on eBay, but most eBay traders who encounter "*informationist"* will not know about my other life as an author and professor.

longer play in the more interesting regions of the game. Other types of communities can use such mechanisms as well. These mechanisms may be explicit, such as requiring a threshold for certain capabilities (e.g., only allowing editors in Wikipedia with more than some number of posts to vote), or they may be implicit, such as providing more weight to old timers (e.g., Digg users with many friends and many front page stories may be more likely to have their stories make it to the front page in the future).

Benefits need not be linked to functionality, either; other factors such as prestige may be effective as well. For example, Slashdot assigns user ids sequentially, with the oldest users having the lowest numbers. Despite conferring no extra privileges, such a mechanism makes early accounts valuable: in 2007 Slashdot included a low user id as one of the items in a charity auction. Someone possessing a low user id would be more willing to accept sanctions to avoid having to start over with a new, high-numbered id.

The benefits of using a long-term pseudonym can be financial as well. In a controlled experiment on eBay, the same seller earned about 8% more revenue selling matched items with an account having high reputation than with new accounts. That means that continuing with an established account is advantageous and thus the threat of sanctions can have a deterrent effect. But the maximum sanction will be equivalent to no more than 8% of the future sales revenue required to build up a new account's reputation (P. Resnick, Zeckhauser, Swanson, & Lockwood, 2006).

Some authors have suggested cross-community reputation systems (e.g., Ba reference) as a way to further increase the value of continued use of a single pseudonym, since people would have to abandon any benefits associated with the existing pseudonym in all the linked communities. But cross-community reputation systems are difficult to implement. Several companies and open source efforts have tried and failed to gain widespread adoption for "open" reputation systems.

Design Claim 28: Increasing the benefits of participating with a long-term identifier will increase the community's ability to sanction misbehavior.

A second way to maintain effective sanctions in the presence of cheap pseudonyms is to make the pseudonyms expensive. For example, in the case that an invitation is necessary to create a new account, the effort involved in gaining an invitation may prevent members from creating an alternative persona and from misbehaving in their primary persona. Another possibility is to charge an entrance fee. Online multiplayer games, like World of Warcraft, that limit the number of player characters enabled for a registered account, use this approach. We have already discussed CAPTCHAS, which impose a small time cost, and thus deter the creation of thousands of accounts, though not the creation of a few.

At the extreme, it may be possible to completely prevent people from getting a second pseudonym once they have acquired one. A person wishing to obtain a pseudonym to participate in a community would have to provide a real-world identity credential to a registration authority, such as a credit card or driver's license. The registration authority would check to make sure that the real-world identity had not previously been issued a pseudonym. Although the community for doctors, Sermo.com, allows members to participate under pseudonyms, it requires them to register with a name that can be checked against national databases of physicians, and that makes

it very difficult to create a second account. Using a cryptographic technique called blind signatures, the registration authority could even be prevented from knowing the mapping from the user's real-world identifier to the user's pseudonym for the community, if that level of anonymity were important to users [cite Friedman and Resnick, 2001].

Design Claim 29: Imposing costs for or preventing pseudonym switching increases the community's ability to sanction misbehavior.

Rather than increasing the costs of creating a new pseudonym, another strategy is to require each new entrant to put something at stake that will be lost if the community decides to sanction the newcomer. One possibility is to require new members to post bonds that will be refunded if they build up good reputations in the community. A related strategy is to tie the reputation of existing members to new members who they invite. Many studies have shown that recruiting new employees in organizations via referrals from existing ones is superior than more formal recruiting methods (e.g., Kirnan, Farley, & Geisinger, 1989). In part this is because referrals lead to better fit, because sponsors know about both the candidate and organization and have an incentive to represent both accurately. The new member has an incentive to make their sponsor look good, and the sponsor has an incentive to help the new member learn the norms and regulations of the community so as to avoid violating them. An additional explanation is that the sponsorship creates incentives for both the sponsor and newcomer to behave well. If the new members misbehave, sanctions may be visited on their sponsoring member.

Design Claim 30: Forcing newcomers to post bonds that may be forfeited if the newcomers misbehave or forcing newcomers' sponsors to stake their own reputations increases the community's ability to sanction misbehavior

Ostrom's fifth principle, culled from studies of successful institutions for managing common pool resources, is the need for graduated sanctions. One reason is that sanctions disproportionate to the offense may be perceived as unfair and illegitimate. She writes, "A large monetary fine imposed on a person facing an unusual problem may produce resentment and unwillingness to conform to the rules in the future." Minor sanctions, proportionate to the offense, are perceived as more legitimate, and errors in their application are also more tolerable. Since the decision about whether to categorize something as deliberate misbehavior versus an accidental or unknowing violation is noisy and subject to biases (e.g., newcomers are more likely to be considered deliberate violators than oldtimers; Hollander, 1971), lighter sanctions mitigate the ill effects from inevitable mistakes in categorization. Stronger sanctions are perceived as more legitimate when applied only after lighter sanctions have proven ineffective.

People tend to be happier and feel they have been treated fairly and with more respect when they are persuaded to comply through expertise and judgment rather than commands and force (Koslowsky et al 2001;Tyler 1997). Sometimes authorities use both nonforceful and forceful measures to gain compliance, and this strategy works as long as the forceful measures do not undercut the persuasiveness of nonforceful measures (Emans et al. 2003). Graduated sanctions that begin with persuasion based on expertise and judgment and proceed to more forceful measures can be especially effective.

In the online community setting, the lowest level of sanctions is a private message explaining the infraction, ideally accompanied by a link to an articulated guideline or rule, and an invitation to discuss the matter further if desired. Unlike more public disapprobation, it allows people to save face. Sanctions can escalate from there, after repeated misbehavior, to public rebuke, disemvoweling or other moderation of individual messages, or gags or bans.

For example, vandals in Wikipedia are initially greeted with an informational message assuming good faith (and allowing the vandal to save face while not continuing to vandalize):

> *Welcome to Wikipedia. Although everyone is welcome to make constructive contributions to Wikipedia, at least one of your recent edits did not appear to be constructive and has been reverted. Please use the sandbox for any test edits you would like to make, and read the welcome page to learn more about contributing constructively to this encyclopedia. Thank you.*

Repeated misbehavior is dealt with through four levels of escalating sanctions and more strongly worded messages, culminating in the brief message:

> *This is the **last warning** you will receive for your disruptive edits. If you vandalize Wikipedia again, you **will** be blocked from editing.*

Design Claim 31: Graduated sanctions increase the legitimacy and thus the effectiveness of sanctions.

Most research on crime prevention shows that perceived certainty of punishment has more deterrence value than factors such as the immediacy or severity of punishment (see Pratt, Cullen, Blevins, Daigle, & Madensen, 2006b for a recent review). A mild but certain punishment is more effective in deterring misbehavior than a severe but uncertain punishment. For instance, the most severe punishment in the U.S. is the death penalty, but this punishment is highly uncertain. The historical evidence is that the death penalty has not deterred murder or rape (see Bailey & Peterson, 1997 for a review; in some cases an increase in these crimes followed executions; Sakamoto, Sekiguchi, Shinkyu, & Okada, 2003). On the other hand, checking the blood alcohol level of every single motorist stopped at sobriety checkpoints is associated with dramatic reductions in drunk driving and alcohol related accidents. And across states, mild versus severe drunk driving penalties does not differentiate drunk driving rates, but certainty of punishment does. It is especially ineffective to ignore misbehavior that negatively affects a community. Rule breaking that goes without punishment encourages copycat offenses and undermines cooperation. Unpunished rule breaking causes even people predisposed to good behavior to cease doing so or exit (E Fehr & Gächter, 2000). This research has encouraged many real world communities to mildly but reliably fine people for visible instances of rule breaking such as pan handling and littering.

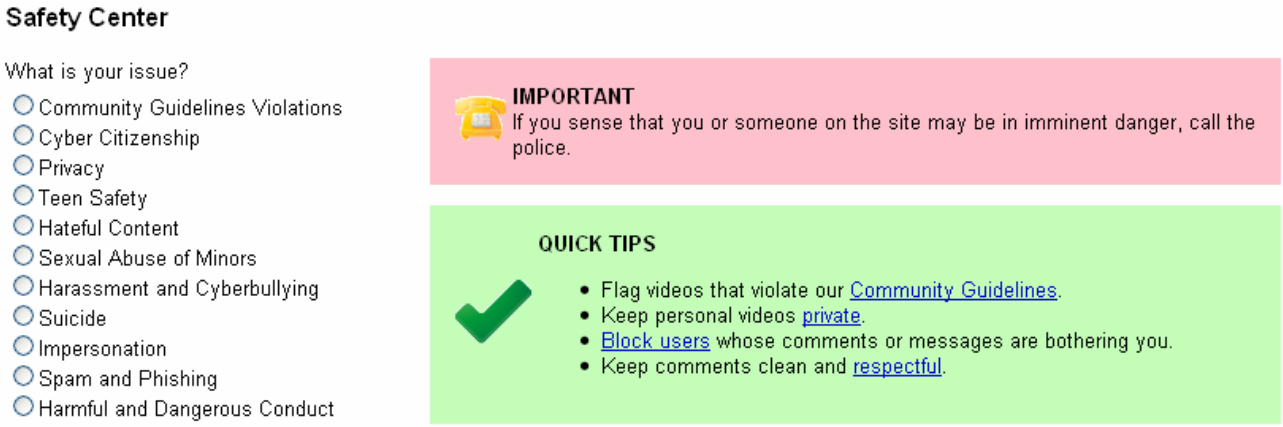


Figure 9. YouTube's Safety Center for reporting inappropriate people or content

One lesson for online communities is that there must be a high probability that norm violations will be detected. One option is that community members can be enlisted to flag violations. For example, YouTube has a safety center, where users can report inappropriate users or content (See Figure 9).Other online communities use software to increase the certainty of detection of inappropriate behavior. Some companies, for example, use software to flag photographs with large flesh-colored areas as potentially pornographic. Facebook uses software to detect when users or applications send requests to too many subscribers and then bans their accounts for a period. In many companies, the flagged material is then handed off to company employees or an outsourcing firm for further evaluation (Stone, 2010).

Design Claim 32: Peer reporting or automatic detection of violations increases the deterrent effect of sanctions.

To enhance certainty of sanctions for violations, there must also be a high probability that sanctions will be imposed after a violation is detected. In many online communities, many of the sanctions will be decided and carried out by members, not external administrators. But the members may not actually impose the sanctions. And for good reason: it is often costly for the person imposing the sanctions. The sanctions may lead to interpersonal drama, and require a significant amount of time and emotional energy for the sanctioning party to defend a decision. Moreover, there may be retaliation against the sanctioning party. There have been instances in which offenders have harassed members who tried to sanction them. On eBay, leaving a negative feedback often led to receiving a negative feedback in return (Chrysanthos Dellarocas & Wood, 2008) . One buyer explained: "I've had a few experiences where I didn't leave non-positive feedback I felt was warranted only to avoid the retaliatory negative (sqpantz, 2008)."

Ostrom and others refer to the delivery of sanctions as a second-order social dilemma or free-rider problem. She quotes Jon Elster, discussing the problem in the context of union members sanctioning (or not) workers who don't join the union (Elster, J. 1989, pp. 41. The Cement of Society. A Study of Social Order. Cambridge University Press):

> *Why, for instance, should a rational, selfish worker ostracize or otherwise punish those who don't join the union? What's in it for him? True, it may be better for all members if all punish non-member than if none do, but for each member it may be even better to remain passive. Punishment almost invariably is costly to the punisher, while the benefits from punishment are diffusely distributed over the members.*

And yet, in many situations, people do voluntarily sanction others, even at some cost to themselves. One common laboratory experiment is called the ultimatum game. One party, the proposer, is given a sum of money. She chooses a division of the money between herself and the decider. If the decider accepts, they each keep the proposed share. If not, neither gets any money. When proposers offer too small a share, many deciders will reject the proposal, punishing the proposer for the unfair proposed division, but at a cost to themselves. In the United States, for example, when offers of a 70%-30% split of the money were made, more than three quarters of deciders rejected the offers. Somewhat fewer accepted bad splits in Slovenia, somewhat more in Japan and Israel. But even in Israel nearly one-third rejected the 70-30 split offers, and two-thirds rejected 90-10 offers (Roth, Prasnikar, Okuno-Fujiwara, & Zamir, 1991).

What can designers do to increase the likelihood that members will impose sanctions when they are warranted? Some of the techniques described earlier also have the desirable side effect of increasing members' willingness to carry out sanctions. First, anything that increases community cohesion will help. Cohesive communities are more likely than non-cohesive ones to both have well defined norms and to enforce them by sanctioning misbehaviors. In particular, much scholarly research suggests that people sanction misbehaviors because in the long run doing so improves the welfare of the groups of which they are apart. Consistent with this logic, Horne showed in a series of experiments that individuals in more cohesive groups (i.e., ones in which individuals are interdependent upon each other) were more likely to enforce norms through social sanctioning (Horne, 2001, 2007; Horne & Cutlip, 2002). Second, graduated sanctions can help. The lowest level, lighter sanctions, tend to be lower in cost to initiate than severe sanctions, which often require significant justification and debate. Third, explicit rules and guidelines that are referenced when applying sanctions can limit the amount of justification and debate that will occur afterward.

Finally, as experiments by Small and Loewenstein's (2005) show, people are more punitive towards identified wrongdoers than toward equivalent, but unidentified, wrongdoers. They propose that identifying an offender increases people's punitiveness because of the stronger feelings people have towards identified others. In support of this thesis, these researchers found that people's anger is much harsher toward identified offenders than unidentified offenders. Thus, when bringing instances of misbehavior to the attention of people deciding on sanctions, identifying the perpetrators by name or picture should increase the willingness to impose sanctions.

There are also some additional measures, not discussed previously, that designers can take. First, the community can designate formal sanctioning roles, so that those imposing sanctions have legitimacy. In message boards or blogs these are typically called moderators; in wikis, administrators. For example, a message with a gentle correction coming from someone who is a

designated moderator is less likely to generate drama or retaliation than the same message coming from someone without a formal role.

Second, steps can be taken to prevent direct retaliation. For example, in 2008 eBay introduced a new rule by which sellers are not allowed to submit negative or neutral feedback anymore, only positive. That eliminated the possibility of sellers retaliating with negative feedback when they received it, and should have made buyers more willing to give negative feedback.

Design Claim 33: Increased community cohesion, graduated sanctions, explicit rules, identifiable perpetrators , formal sanctioning roles, and anti-retaliatory measures increase the likelihood that sanctions will be applied and thus increase the deterrent effect of sanctions.

Summary of Design Alternatives

This chapter has explored means for regulating behavior that violates behavioral norms, both limiting it and limiting the damage that it causes when it does occur. We conclude with a summary of the design alternatives considered throughout the chapter.

Several options are available that alter how information is used or displayed. Inappropriate posts can be moved to areas where they are less likely to be seen, scored so that they will be hidden from other users, or degraded through techniques like disemvoweling. Bad edits in wikis can be reverted, making them invisible except to people who examine a page's history. Ratings that are suspected to be from manipulators can be removed or discounted when making recommendations. Links can be annotated so that search engines will ignore them. These options can limit the damage that non-normative behavior causes and/or can reduce incentives for doing such behavior in the first place.

Feedback and rewards can come in other forms as well. Feedback may be directly solicited and displayed along with messages, or it may be aggregated into reputation profiles. A good profile can lead to rewards in the form of better treatment from members in the future and the reverse for bad reputations. Instead of affecting reputation profiles, feedback about individual actions can lead to monetary payoffs (positive or negative) or payoffs in an internal currency that has value within the community. Or rewards or sanctions can be assigned to someone's cumulative behavior: a bond posted upon entry into the community can be forfeited if the person misbehaves.

Several technical features can be used to limit the actions available to people who may violate behavior norms. Throttles or activity quotas can limit repetitive behavior. Charging for actions using a currency accrued through normal participation can also limit repetitive behavior. It also For members who gain value from normal participation, such charges do not create a binding constraint but they serve as a disincentive for trolls and attackers, because earning the currency may be costly for them. Gags and bans can silence bad actors altogether. To prevent people from sidestepping gags and bans, or any form of sanction, the account registration process can impose limits or costs on the creation of new accounts.

Roles, rules, policies, and procedures play a big part in regulating non-normative behavior. Having clear rules and policies, and having fair procedures for applying any of the filters, sanctions, and participation limits, will decrease resistance to them. Legitimacy will also increase and resistance decrease if there is wide participation in setting of rules and policies, and if the enforcement roles are widely distributed. Two particular features of the contents of rules and policies are also helpful. Sanctions should be graduated, both to increase their legitimacy and to increase the willingness of enforcers to apply them. And everyone should learn to ignore trolls.

Finally, as in other chapters, we find that there is considerable power in decisions about framing, ways of communicating what is happening in the community. Highlighting or leaving traces of bad behavior and prominently displaying behavior guidelines can help to clarify norms, but runs the risk of conveying a descriptive norm that misbehavior is rampant. Showing names and

pictures of those who took actions will make people think twice about misbehaving and increase people willingness to enforce sanctions against those who do misbehave. Framing disciplinary actions in a way that allows people to save face ("Someone using your account…", or "You may not have realized…") can make people more receptive and willing to change their behavior.

In the face of harmful behavior, it may feel natural to turn first to tangible remedies, such as removing bad posts or banning or throttling the posters. An important theme of the chapter is that less tangible, softer and more behavioral remedies may be desirable to try first. Guidelines can be clarified, and the community as a whole can be involved in that process, in order to build legitimacy. Individuals can be reminded and corrected in a way that allows them to save face. Off-topic communication can be gently encouraged to move to an interaction space where people don't mind the digressions. Trolls can be ignored. Responses can escalate if these mild approaches fail, with other behavioral remedies such as public rebuke. Behavioral responses, however, will not always be sufficient. Especially in the face of manipulators and spammers who create lots of accounts and act through bots, or in the face of trolls who gain rather than lose utility from other members getting mad at them, communities will need some more automated and tangible ways to limit damage.

Type	**Design Alternative**	**Claim #**
Selection, sorting, highlighting		
	Moderation systems that pre-screen, degrade, label, move, or remove inappropriate messages	Design Claim 1
	Redirecting inappropriate posts to other places	Design Claim 2
	Reversion tools	Design Claim 5
	Filters or influence limits	Design Claim 6
	Telling search engines not to follow links	Design Claim 24
Community structure		
	cohesive groups	Design Claim 21 Design Claim 33
Feedback and Rewards		
	Displaying feedback of members to others	Design Claim 16
	Reputation systems	Design Claim 26
	Prices, bonds, or bets that make undesirable actions more costly than desirable actions	Design Claim 27
	Increasing the benefits of participating with a long-term identifier	Design Claim 28
	Forcing newcomers to post bonds that may be forfeited if the newcomers misbehave or forcing newcomers' sponsors to stake their own reputations	Design Claim 30
Access Controls		
	Activity quotas	Design Claim 8
	Gags and bans	Design Claim 9
	Paying to take actions in the community with currency accumulated through normal participation	Design Claim 11

	Limiting fake accounts with CAPTCHAs or identity checks	Design Claim 12
	Imposing costs for or preventing pseudonym switching	Design Claim 29
Roles, rules, policies, and procedures		
	Consistently applied moderation criteria, a chance to argue one's case, and appeal procedures	Design Claim 3
	Moderation decided by people who are members of the community, are impartial, and have limited or rotating power	Design Claim 4
	A widely followed norm of ignoring trolls	Design Claim 7
	Consistently applied criteria for gags and bans	Design Claim 10
	Explicit rules and guidelines	Design Claim 18 Design Claim 19 Design Claim 33
	Community influence on rule making	Design Claim 22
	Graduated sanctions	Design Claim 31 Design Claim 33
	Peer reporting or automatic detection of violations	Design Claim 32
	Formal sanctioning roles	Design Claim 33
	Anti-retaliatory measures	Design Claim 33
Presentation and Framing		
	Publicly displaying examples of appropriate behavior	Design Claim 13
	Publicly contrasting examples of inappropriate behavior in the context of a descriptive norm of appropriate behavior	Design Claim 14
	Publicly displaying many examples of inappropriate behavior	Design Claim 15
	displaying statistics that highlight the prevalence of normative behavior	Design Claim 17
	Prominently displayed guidelines	Design Claim 19
	reminders at the point of an action that may violate norms	Design Claim 20
	Face-saving ways to correct norm-violations	Design Claim 23
	Verified identities and pictures	Design Claim 25 Design Claim 33

REFERENCES

47jamoo (2007, May 18). Definition Of God-Modding Retrieved Oct 4, 2010, from p://marvelheroesrpg.proboards.com/index.cgi?board=rules&action=display&thread=8

Akdeniz, Y. (2008). *Internet child pornography and the law: national and international responses.* . Farnham, Surrey, UK: Ashgate Publishing, Ltd.

Alpizar, F., Carlsson, F., & Johansson-Stenman, O. (2008). Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. *Journal of Public Economics, 92*(5-6), 1047-1060.

Axelrod, R., & Keohane, R. (1985). Achieving cooperation under anarchy: Strategies and institutions. *World Politics: A Quarterly Journal of International Relations, 38*(1), 226-254.

Bailey, W., & Peterson, R. (1997). Murder, capital punishment, and deterrence: A review of the literature. In H. A. Bedau (Ed.), *The death penalty in America: Current controversies* (pp. 135-161). New York: Oxford University Press.

Bajaj, V. (2010, April 25). Spammers pay others to answer security tests. *New York Times,* p. B6,

Bajari, P., & Hortacsu, A. (2004). Economic insights from internet auctions. *Journal of Economic Literature, 42*(2), 457-486.

Bartle, R. (1996, April). Hearts, Clubs, Diamonds, Spades: Players Who Suit Mud Retrieved Sept 18, 2007, from http://www.mud.co.uk/richard/hcds.htm

Bordia, P. (1997). Face-to-face versus computer-mediated communication: A synthesis of the experimental literature. *Journal of Business Communication, 34*(1), 99.

bunkerbuster1 (2006, June 8). What is a concern troll Retrieved Oct 4, 2010, from http://www.democraticunderground.com/discuss/duboard.php?az=view_all&address=364x1382185

Burke, M., Kraut, R. E., & Williams, D. (2010). Social Use of Computer-Mediated Communication by Adults on the Autism Spectrum *CSCW'2010: Proceedings of the ACM Conference on Computer Supported Coooperative Work* (pp. 425-434). NY: ACM Press.

Cabral, L., & Hortacsu, A. The dynamics of seller reputation: Evidence from eBay. *The Journal of Industrial Economics, 58*(1), 54-78.

Cialdini, R. (2003). Crafting normative messages to protect the environment. *Current Directions in Psychological Science, 12*(4), 105.

Cialdini, R., Kallgren, C., & Reno, R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in experimental social psychology, 24*(20), 1-243.

CmdrTaco (2003, JUne 4). How does moderation work? Retrieved Oct 4, 2010, from http://slashdot.org/faq/com-mod.shtml

Colquitt, J., Conlon, D., Wesson, M., Porter, C., & Ng, K. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology, 86*(3), 425-445.

Cornwell, B., & Lundgren, D. (2001). Love on the Internet: Involvement and misrepresentation in romantic relationships in cyberspace vs. realspace. *Computers in Human Behavior, 17*(2), 197-211.

Crystal (2009, June 14). The Twitter Rules Retrieved Oct 4, 2010, from http://twitter.zendesk.com/forums/26257/entries/18311

DeJong, W., Schneider, S., Towvim, L., Murphy, M., Doerr, E., Simonsen, N., et al. (2006). A multisite randomized trial of social norms marketing campaigns to reduce college student drinking. *Journal of Studies on Alcohol, 67*(6), 868.

Dellarocas, C., & Wood, C. (2006). *The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias*. Research Paper Research Paper Robert H. Smith School University of Maryland College Park, Maryland (No. RHS 06-041) http://ssrn.com/abstract=923823

Dellarocas, C., & Wood, C. A. (2008). The Sound Of Silence In Online Feedback: Estimating Trading Risks In The Presence Of Reporting Bias. *Management Science., 54*(3), 460-476.

Dewally, M., & Ederington, L. (2006). A comparison of reputation, certification, warranties, and disclosure as remedies for information asymmetries: Lessons from the on-line comic book market. *Journal of Business, 79*(4).

Dibbell, J. (1993, Dec 23). A rape in cyberspace: How an evil clown, a haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *The Village Voice*, from http://www.juliandibbell.com/texts/bungle_vv.html

Donath, J. S. (1997). Identity and deception in the virtual community. In P. Kollock & M. Smith (Eds.), *Communities in Cyberspace.* (pp. 2-59). Berkeley:: University of California Press.

Ducheneaut, N. (2005). Socialization in an Open Source Software Community: A Socio-Technical Analysis. *Computer Supported Cooperative Work, 14*(4), 323 - 368

Durkheim, E. (1953 [1903]). The determination of moral facts (D. F. Pocock, Trans.) *Sociology and philosophy* (pp. 35-63). London: Cohen and West.

Facebook (2010). Warnings Retrieved Oct 4, 2010, from http://www.facebook.com/help/?page=421

Fehr, E., Fischbacher, U., & Gächter, S. (2003). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature, 13*(1), 1-25.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90*(4), 980-994.

Festinger, L., Pepitone, A., & Newcomb, T. (1952). Some consequences of de-individuation in a group. *Journal of Abnormal and Social Psychology, 47*(2), 382-389.

Gibbs, J. P. (1985). Deterrence theory and research. *Nebraska Symposium on Motivation, 33*, 87-130.

Goffman, E. (1959). *The presentation of self in everyday life*. Garden City, NY: Doubleday.

Goffman, E. (1963). *Behavior in public places*. New York: The Free Press.

Halfaker, A., Kittur, A., Kraut, R., & Riedl, J. (2009). A Jury of Your Peers: Quality, Experience and Ownership in Wikipedia *WikiSym 2009: Proceedings of the 5th International Symposium on Wikis and Open Collaboration.* . New York: : ACM Press.

Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. (2002). Searching for safety online: Managing" trolling" in a feminist forum. *The Information Society, 18*(5), 371-384.

Hoag, D. (1996, Jaunary). The Well Host Manual 4.4. Retrieved August 11, 2003, from http://www.well.com/~confteam/hostmanual/

Horne, C. (2001). The enforcement of norms: Group cohesion and meta-norms. *Social Psychology Quarterly, 64*(3), 253-266.

Horne, C. (2007). Explaining norm enforcement. *Rationality and Society, 19*(2), 139.

Horne, C., & Cutlip, A. (2002). Sanctioning costs and norm enforcement: An experimental test. *Rationality and Society, 14*(3), 285-307.

Jackson, G. A. (1994). Promoting Network Civility At MIT: Crime & Punishment, Or The Golden Rule? Retrieved October 4, 2010, from http://www.mit.edu/activities/safe/data/mit-stopit.html

Joeblo Movie Club (2005). Basic rules and guidelnes Retrieved Dec 6, 2005, from http://www.joblo.com/forums/announcement.php?s=641f9cd5b47beab4ad423f0c861dba3c&forumid=21

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality & Social Psychology, 65*(4), 681-706.

Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*(5), 754-775.

Kiesler, S., Zubrow, D., Moses, A., & Geller, V. (1985). Affect in computer-mediated communication: An experiment in synchronous terminal-to-terminal discussion. *Human-Computer Interaction, 1optional*, 77-104.

Kirnan, J., Farley, J., & Geisinger, K. (1989). The relationship between recruiting source, applicant quality, and hire performance: An analysis by sex, ethnicity, and age. *Personnel psychology, 42*(2), 293-308.

Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. (2007). He says, she says: Conflict and coordination in Wikipedia *CHI 07: Proceedings of the ACM Conference on Human Factors in Computing Systems* (pp. 453-462). New York, NY: ACM Press

Kollock, P., & Smith, M. (1996). Managing the virtual commons: Cooperation and conflict in computer communities. In S. Herring (Ed.), *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives* (pp. pp. 109-128). Amsterdam: John Benjamin.

Krogh, G. v., Spaeth, S., Lakhani., K. R., & Hippel, E. v. (2003). Community, Joining, and Specialization in Open Source Software Innovation: A Case Study. *Research Policy, 32*(7), 1217-1241.

Lessig, L. (1999). *Code and other laws of cyberspace*: Basic books.

Leventhal, G. (1976). The Distribution of Rewards and Resources in Groups and Organizations1. In L. Berkowitz & W. Walster (Eds.), *Advances in experimental social psychology* (Vol. 9, pp. 91-131). New York: Academic Press.

Lewis, S., Langan, C., & Hollander, E. (1972). Expectation of future interaction and the choice of less desirable alternatives in conformity. *Sociometry, 35*(3), 440-447.

Mobasher, B., Burke, R., Bhaumik, R., & Williams, C. (2007). Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Transactions on Internet Technology (TOIT), 7*(4), 23.

Moon, J., & Sproull, L. (2008). The role of feedback in managing the Internet-based volunteer work force. *Information Systems Research, 19*(4), 494-515.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*: Cambridge University Press.

Ostrom, E. (2000). Collective action and the evolution of social norms. *The Journal of Economic Perspectives, 14*(3), 137-158.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). *The pagerank citation ranking: Bringing order to the web*. Stanford, CA: Stanford University.

Perkins, H., Meilman, P., Leichliter, J., Cashin, J., & Presley, C. (1999). Misperceptions of the norms for the frequency of alcohol and other drug use on college campuses. *Journal of American College Health, 47*(6), 253-258.

Postmes, T., & Spears, R. (1998). Deindividuation and antinormative behavior: A meta-analysis. *Psychological Bulletin, 123*(3), 238-259.

Pratt, T., Cullen, F., Blevins, K., Daigle, L., & Madensen, T. (2006a). The empirical status of deterrence theory: A meta-analysis. *Taking stock: The status of criminological theory, 15*.

Pratt, T., Cullen, F., Blevins, K., Daigle, L., & Madensen, T. (2006b). The empirical status of deterrence theory: A meta-analysis. In F. Cullen, J. Wright & K. Blevins (Eds.), *Taking stock: The status of criminological theory* (Vol. 15, pp. 367-396). New Brunswick NJ: Transaction Publishers.

PsychCentral (2008, Mar 5). Terms of Use, from http://psychcentral.com/about/terms.htm

Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM, 43*(12), 45-48.

Resnick, P., & Sami, R. (2007). The influence limiter: provably manipulation-resistant recommender systems *Proceedings of the 2007 ACM Conference on Recommender Systems* (pp. 25-32). New York: ACM.

Resnick, P., & Sami, R. (2008). The information cost of manipulation-resistance in recommender systems *Proceedings of the 2008 ACM Conference on Recommender Systems* (pp. 147-154). New York: ACM.

Resnick, P., & Zeckhauser, R. (2002). Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. *Advances in Applied Microeconomics: A Research Annual, 11*, 127-157.

Resnick, P., Zeckhauser, R., Swanson, J., & Lockwood, K. (2006). The value of reputation on eBay: A controlled experiment. *Experimental Economics, 9*(2), 79-101.

Rohrer, J., Baron, S., Hoffman, E., & Swander, D. (1954). The stability of autokinetic judgments. *Journal of Abnormal and Social Psychology, 49*(4 Pt 1), 595-597.

Rosenberg, M. J. (2009). The conditions and consequences of evaluation apprehension *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books* (pp. 211-263). New York: Oxford University Press.

Roth, A., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *The American Economic Review, 81*(5), 1068-1095.

Sakamoto, A., Sekiguchi, K., Shinkyu, A., & Okada, Y. (2003). Does Media Coverage of Capital Punishment Have a Deterrent Effect on the Occurrence of Brutal Crimes? An Analysis of Japanese Time-Series Data from 1959 to 1990. *Progress in Asian social psychology: conceptual and empirical contributions*, 277.

Sanger, L. (2005). The Early History of Nupedia and Wikipedia: A Memoir. In C. DiBona, D. Cooper & M. Stone (Eds.), *Open Sources 2.0: The Continuing Evolution*. Sebastopol, CA: O'Reilly Media, Inc.

Seay, A. F., & Kraut, R. E. (2007). Project massive: self-regulation and problematic use of online gaming. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 829-838.

Shankland, S. (2000, September 15). New email software can help you bite your tongue. *CNET News*, from http://news.cnet.com/New-email-software-can-help-you-bite-your-tongue/2100-1040_3-245790.html#ixzz11PrpARzA

Small, D., & Loewenstein, G. (2005). The devil you know: The effects of identifiability on punitiveness. *Journal of Behavioral Decision Making, 18*(5), 311–318.

Smith, J. (2007). Tragedies of the ludic commons-understanding cooperation in multiplayer games. *Game Studies, 7*(1).

Soetevent, A. (2005). Anonymity in giving in a natural context--a field experiment in 30 churches. *Journal of Public Economics, 89*(11-12), 2301-2323.

sqpantz (2008). Works for me Retrieved Oct 4, 2010, from http://www.techdirt.com/articles/20080205/160733184.shtml

Stone, B. (2010, July 19). Policing the web's lurid precints. *New York Times*, p. B1,

Van Alstyne, M. (2007). Curing spam: rights, signals & screens. *Economists' Voice, 4*(2), Article 4.

Vesperman, J., & Henson, V. (2004). Building and Maintaining an International Volunteer Linux Community *Proceedings of the FREENIX Track: 2004USENIX Annual Technical Conference*. Berkeley, CA: USENIX Association.

Viegas, F., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. *CHI 2004: ACM Conference on Human-Factors in Computing Systems*. NY: ACM Press.

Walker, J., Gardner, R., & Ostrom, E. (1990). Rent dissipation in a limited-access common-pool resource: Experimental evidence. *Journal of Environmental Economics and Management, 19*(3), 203-211.

Wikipedia (2010a, Sept 22). Spam in blogs Retrieved Oct 4, 2010, from http://en.wikipedia.org/wiki/Spam_in_blogs

Wikipedia (2010b, Sep 20). Wikipedia:Blocking policy Retrieved Oct 4, 2010, from http://en.wikipedia.org/wiki/Wikipedia:Blocking_policy

Wikipedia (2010c, Aug 22). Wikipedia:WikiProject Vandalism studies/Study1 Retrieved Oct 4, 2010, from http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Vandalism_studies/Study1

Williams, G. A. (1997). Hosting: Online Moderator Guidelines and Community-Building Tips Retrieved Oct 4, 2010, from http://www.well.com/confteam/hosting.html

Wookiepedia (2010, Apr 5). Wookiepedia: The Star Wars encyclopedia that anyone can edit Retrieved Oct 4, 2010, from http://starwars.wikia.com/wiki/Main_Page

Zitek, E., & Hebl, M. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology, 43*(6), 867-876.