Human Computation (2017) 4:1:78–106

ISSN: 2330-8001, DOI: 10.15346/hc.v4i1.5

Can you moderate an unreadable message? "Blind" content moderation via human computation

SETH FREY, University of California, Davis

MAARTEN W. BOS, Disney Research

ROBERT W. SUMNER, Disney Research

ABSTRACT

User-generated content (UGC) is fundamental to online social engagement, but eliciting and managing it come with many challenges. The special features of UGC moderation highlight many of the general challenges of human computation in general. They also emphasize how moderation and privacy interact: people have rights to both privacy and safety online, but it is difficult to provide one without violating the other; scanning a user's inbox for potentially malicious messages seems to imply access to all safe ones as well. Are privacy and safety opposed, or is it possible to guarantee the safety of anonymous content without access to that content? We demonstrate that such "blind content moderation" is possible in certain circumstances. Additionally, the methods we introduce offer safety guarantees, an expressive content space, and require no human moderation load: they are safe, expressive, and scalable. Though it may seem impossible to try moderating UGC without human- or machine-level access to it, human computation makes blind moderation possible. We establish this existence claim by defining two very different human-computational methods, behavioral thresholding and reverse correlation. Each leverages the statistical and behavioral properties of so-called inappropriate content in different decision settings to moderate UGC without access to a message's meaning or intention. The first, behavioral thresholding, is shown to generalize the well-known ESP game.

1. INTRODUCTION

Ubiquitous user-generated content (UGC) is one of the most notable features of an Internet-driven culture. Even ignoring its obvious benefits to platforms such as YouTube that are driven

almost entirely by UGC, such content can foster in users a sense of ownership and creative engagement with others (Ekstrom & Ostman, 2015; Subramaniam, Valdivia, Pellicone & Neigh, 2014). However, the vast freedom UGC provides to creators makes it inherently difficult to govern. Norm-violating content is a sad fact of online social systems, one with temporal, monetary, and even human costs (Cheng, Danescu-Niculescu-Mizil & Leskovec, 2015; Grimmelmann, 2015; Harrison, 2010; Hermida & Thurman, 2007; Stefanovitch, Alshamsi, Cebrian & Rahwan, 2014). Malicious users can attack or endanger others with unsafe content, and techniques for preventing such attacks face a variety of challenges. Manual filtering can be prohibitively expensive at scale, while algorithmic filtering methods, although more likely to scale, are limited in effectiveness by the current states of the arts of NLP, AI, and machine learning (Hidalgo, Sanz, García & De Buenaga Rodríguez, 2009; Sood, Antin & Churchill, 2012).

Institutions that can more efficiently and reliably flag inappropriate content can not only better leverage the benefits of collective action online; they can also make it safer and easier to be inclusive of users of different ages and backgrounds (Subramaniam et al., 2014). But inclusiveness introduces its own complications. UGC-moderation settings seem to put privacy and safety at odds: Imagine that user Alice sends a message to little user Bobby through Charlie's communication system. Charlie can offer Bobby safety guarantees, perhaps by reviewing Alice's content and intercepting messages from her that violate external norms of propriety. Charlie can also offer Bobby privacy guarantees by precluding his own system's access to the actual content, perhaps by providing Alice with cryptographic tools within his system. But can Charlie simultaneously guarantee both Bobby's safety and privacy? In the context of a public-key cryptographic system, providing safety would require access to the private key, while providing privacy would proscribe such access. So what kind of privacy can be offered within a system that also has moderation obligations (Mont, Pearson & Bramhall, 2003)? Are privacy of content and protection from content mutually exclusive qualities?

Whether due to the requirements of the audience or the demands of the law, UGC systems are facing increasingly clear pressures to provide for both the privacy and safety of their audiences (Barnes, 2006; Belenkiy et al., 2007; Silverstein, Nissenbaum, Flanagan & Freier, 2006). The technical challenges in reconciling these and other conflicting demands on UGC moderation are as interesting practically as they are theoretically. Problems of UGC management and moderation are problems of designing constraints to improve collective outputs in social context. They are particularly important problems because of how deeply UGC is embedded in peoples' lives on the Internet. For these reasons, research on UGC management is a strategic source of advances for the human computation community.

1.1 Goal

Our main goal is to showcase "blind" UGC moderation: that which guarantees both privacy and safety via content moderation that is blind to content semantics. We propose that interaction designers can accomplish blindness by leveraging the statistical structure of inappropriate content, regularities in human collective behavior, and the social and cognitive traits of specific kinds of user communities. The main contribution of this work is the specification of two moderation methods that accomplish this goal with the help of constraints specific to two more narrow problem domains, namely child-focused UGC and "collective" UGC. Though we present our methods as theoretical constructs that establish the non-impossibility of blind moderation, they are in principle simple and flexible enough to be applied to diverse UGC media including text, chat, and video content. Although they can be combined with other established methods to increase the effectiveness of existing UGC platforms, we do not present these methods as production-ready systems, but as stand-alone existence proofs of our unlikely claim that unobservable content can, in certain contexts, still be moderated.

1.2 Literature

Human computation shows that—with clever incentives and constraints—human collectives can be made to behave like algorithms (Das & Vukovic, 2011; Jain & Parkes, 2009; Yuen, Chen & King, 2009). This is as true for UGC moderation as it is for other areas of human computation. Recognizing that human moderators are costly, some human computation researchers have looked to crowdsourcing to increase the scalability of professional content moderation (Ghosh, Kale & McAfee, 2011). Refinements of this strategy have used automatic reputation calculations to identify and elevate the most reliable crowd workers (Adler & de Alfaro, 2007; Pantola, Pancho-Festin & Salvador, 2011). Ghosh and McAfee take a game-theoretic approach, reducing the need for content moderation with mechanisms that remove users' incentives to emit low-quality UGC in the first place (Ghosh & McAfee, 2011). Still other projects have focused on identifying those most likely to violate norms of propriety, either via crowdsourcing or through collaborative filtering (Adler & de Alfaro, 2007; Cheng et al., 2015; Wang et al., 2012). Despite the success of these approaches in providing safety, they all do so at the expense of privacy. In other words, none of these approaches is blind.

Privacy and security researchers have attended extensively to the problem of providing privacy in the kinds of social media environments that UGC represents (van den Berg, Pötzsch, Leenes, Borcea-Pfitzmann & Beato, 2011), and they have long recognized the tension between respecting the privacy of users and holding them accountable for their behavior (Belenkiy et al., 2007; Burmester, Desmedt, Wright & Yasinsac, 2006; Mont et al., 2003; Pearson et al., 2009). Research on accountable privacy in online social media applications has restricted itself to solutions based on access control and authentication (Gates, 2007; Pang & Zhang, 2015; Sayaf & Clarke, 2013; Shehab, Squicciarini, Ahn & Kokkinou, 2012), although human-computational approaches have also attracted interest (Squicciarini, Shehab & Paci, 2009; Squicciarini, Shehab & Wede, 2010).

Digital-forensics researchers have also encountered the conflict between moderating content and forgoing access to it. A number of legal and ethical issues surround the problem of detecting and reporting child pornography, because of the blurry lines between storing it and owning it, detecting it and consuming it, and reporting it and propagating it (McIntyre, 2012). In early solutions, the problem of detection was left to the National Center for Missing and Exploited Children, who then distributed hash sums of exploitative image files to major social media providers such as Microsoft, Facebook, Google, and Twitter. However, the effectiveness of this approach was undermined by that fact that hashes change completely with arbitrarily small changes to their inputs. Modern solutions have approached this problem with fuzzy, "robust," or locality-sensitive hashes that are insensitive to small changes to an image (Bjelland, Franke & Årnes, 2014). Within this system, UGC moderation teams are able to filter out unsafe media without any access to the semantics of that content.

More generally, contemporary privacy research continues to be firmly grounded in a paradigm in which providing privacy means providing access to technical schemes that offer users control over information about them (Levin & Abril, 2008). The weakness of this perspective is that it is more prone to treating concepts such as privacy and safety as culturally static; the tracking of loyalty to a grocery store is governed by different cultural standards of propriety than the tracking of something like loyalty to a government. Alternatively, Nissenbaum's paradigm of contextual privacy argues that considerations of context should be at the center of the privacy-design process (Nissenbaum, 2004; 2009). These concerns are particularly important for sensitive populations, such as young people (Barnes, 2006; Silverstein et al., 2006). Human-computational approaches are an important frontier in privacy research because humans naturally take social context into account, and the human-computed outputs of norm-respecting community members can be leveraged to tacitly integrate social context in collective outputs. This ability of human computation to naturally accommodate the complexities of social context will be evident in our definition of "inappropriate content," below.

2. CONCEPTUAL FRAMEWORK

Our main approach is based on a statistical perspective on UGC management. But before introducing our methods, we will motivate them conceptually with a large (but non-comprehensive) survey of production UGC systems. This discussion will not only give a sense of the range of approaches to content moderation; it will allow us to elaborate on the ideas of safety and privacy, to add two more features common to UGC moderation systems, and to speculate as to how these four features are traded off in different approaches to moderation.

2.1 Definitions

UGC platforms, because they control their software, can define moderation in terms of their own norms of propriety. So what makes content inappropriate? It has long been recognized that

communities are capable of developing social norms that are well adapted to their specific form of peer production (Kollock & Smith, 1996; Kraut & Resnick, 2012; Ostrom, 2005). So rather than "hard-coding" social norms, human computation can leverage small communities' abilities to develop local norms. Within this broad scope, "inappropriate" may include content ranging from impolite to illegal. It may also include criteria that do not relate to social propriety at all, such as content that violates a host's technical terms of use or brand guidelines. By designing environments that leave the definitions of "safe" and "appropriate" to the people subject to them, we offer general tools that empower communities to develop their own norms.

Because we appreciate politeness as well as brevity, we will, whenever possible, refer to instances of "inappropriate user-generated content" as "**bleeps**" or "**#@%!s.**"

2.2 Properties of UGC-moderation systems

UGC-moderation challenges have been implicated in the shuttering of at least one major social game (Purslow, 2015), and they are apparent in the design decisions of every social media platform. Industry coinages such as "TTP," referring to the minimum time necessary to produce problematic content within a given UGC system, attest to the immediacy of content-moderation problems in today's social media platforms (Kelly, 2009).

To structure the variety of challenges that UGC systems face, we identify trade-offs between the properties of "safety," "privacy," "expressivity," and "scalability." The safety of a system is its ability to shield message receivers from inappropriate content. By "privacy," we refer narrowly to the ability or inability of a system owner to access items of user content. Expressivity is the ability of a system to offer unrestricted, flexible content-creation tools to all users. Scalability is the ability of a system to grow its user base with a less-than-proportional increase in the quality and costs of moderation. Scalability fails when the need for acts of costly moderation scales linearly or superlinearly with the quantity of content. A symptom of scaling failures is that quality moderation becomes an unsustainable cost center and bottleneck to growth.

2.3 Interactions between properties of UGC-moderation systems

The concepts of expressivity, safety, scalability and privacy capture the tension faced by a variety of existing UGC-moderation systems. Since blindness is not a property of any of the systems we review, this discussion focuses instead on how the three qualities beside privacy are traded off.

For example, there is a direct tension between providing users with expressivity and safety. A system that is both safe and expressive would allow users unlimited freedom to create, up to the point that any particular piece of content violates community norms. But a content moderator (or automated moderation system) necessarily faces two types of content: that which should be censored and that which should be allowed. And that moderator has two corresponding choices: to censor an item of content or let it through. Within this frame, a moderator can err in two ways:

either by allowing content that should have been filtered or by censoring content that should have been allowed. Moderation systems are vulnerable to false negatives (passing inappropriate content) and false positives (filtering appropriate content), and, observed within the framework of signal-detection theory, decreasing the quality of a moderation system is equivalent to increasing at least one of those types of errors. All else held equal, to increase the false-negative rate is to decrease the safety of a system, and to increase the false-positive rate is to decrease the expressivity of a system.

Moving from the pairwise trade-offs between safety and expressivity, in the remainder of this analysis, we explore three-way interactions between safety, expressivity, and scalability. The interactions we explore all suggest that a compromise on one of those qualities is necessary to leave the other two intact. For example, we argue that systems that seem successfully to provide both safety and expressivity, such as child-focused UGC platforms, must compromise on scalability, perhaps by relying on staffs of human moderators, and that systems that seem expressive and scalable, such as Facebook, compromise on safety by requiring users to personally identify themselves. We also discuss trade-offs between safety and the fourth feature, privacy. All of the systems we discuss seem to take for granted that providing safety requires the violation of privacy implied by human- or machine-level access to UGC.

One strategy for maximizing safety and scalability is to compromise on expressivity. Many platforms that are not considered UGC platforms offer superficial customization options, such as color customization, although users may circumvent these implicit controls in unexpected ways (McWhertor, 2008). In the area of text UGC, online news providers that have customarily supported reader commenting are increasingly choosing the safe and scalable route of disabling free-form user contributions (Hughey & Daniels, 2013), in favor of more limited forms of interaction, such as "liking," sharing, and voting. Constraining expressivity, whatever its downsides, has the merit that it can guarantee safety without a large, costly staff of human moderators.

Similarly, designers of social UGC platforms who desire an expressive and scalable system can choose to compromise on safety. The most extreme way to reduce the safety of a UGC system—its ability to expose users to inappropriate content—is to forbid censorship and moderation. A notable example of a site with minimal rules is 4chan.org, which has a famously ebullient and toxic community. A more moderate and common compromise on safety is to forbid anonymity and force contributors to a UGC platform to explicitly invoke their real-world identities. The most notable example is Facebook's "real names" policy, which imposes on users this minimum level of accountability for the content they post. In the words of Facebook's CEO: "We know that people are much less likely to try to act abusively towards other members of our community when they're using their real names" (Davidson, 2015). Relying on the self-policing that comes with personal accountability has two effects. Because self-censorship reduces the demands on moderation staff, scalability remains high. And because the greater proportion of censorship is self-censorship, safety can be maintained with cultural norms instead of artificial technical

constraints that limit expressivity. In the same article, Facebook presents the "real names" policy as increasing rather than decreasing the safety of those who share their names. However, "real names" policies violate the personal safety inherent in being able to remain anonymous. Furthermore, as is clear from Facebook's 13+ age limit, real-name policies are not an appropriate solution for all populations.

Other systems compromise on scalability in ways that seem to benefit their ability to serve safe and expressive UGC. The most straightforward strategy is to employ professional human content moderators (Purslow, 2015). Systems that rely on professional moderators are very costly to maintain to the point that, with a high enough volume of UGC, human-moderation systems may not be viable. Professional moderators in large-scale UGC systems have finite time and resources with which to monitor high-density streams of content. They often sample only a subset of all content, which, by increasing the miss rate, can be seen as trading safety for scalability. And there are other downsides to this kind of strategy. According to a recent *Wired* report, most large-scale social media systems rely on veritable armies of low-wage outsourced workers for the monitoring of user-generated content (Chen, 2014). This report emphasizes the human cost of UGC management strategies that rely upon professional moderators: moderators are exposed to enough violent and illegal content that many employers either offer or require regular psychological reviews.

The trade-offs between different styles of content moderation are rarely so clean-cut as in the examples above. It is much more common for real-world production systems to combine multiple moderation methods in ways that trade off more sensitively between the competing objectives of safety, expressivity, scalability, and privacy on an effective UGC system. Most production platforms rely to some extent on a combination of strategies. These can be designed to integrate the advantages of staffs of professional moderators, conventional computational filtering approaches to content moderation (such as natural language processing), and design decisions that support norms of civility. This will be clear in a discussion below of a proprietary hybrid system, which we call SafetyText, and on which we validate the first "blind" moderation method, behavioral thresholding.

3. BLIND UGC MODERATION

Is it possible to moderate user-generated content without having access to it? We present two blind moderation methods: behavioral thresholding and reverse correlation (BT & RC). Both rely on the assumption that norm-violating content is, by definition, less common than abiding content. In behavioral thresholding, input content is passed if it is sufficiently common while, in reverse correlation, multiple items of input content are aggregated to produce a single "collective" UGC output that averages out rare deviations. BT is blind because it can work on cryptographic hashes, while RC is blind in the more limited sense that the semantics of any individual item of content are so unclear as to be nearly opaque.

We present these methods as existence proofs rather than as proposed systems. Despite this more theoretical focus, we do present an empirical validation of BT, and we describe RC in terms of previous studies that have established its effectiveness in domains outside of human computation. Because we do not want a proprietary game, a specific system, or implementation details to distract from our core existence claim, we treat both methods in generic terms, with both general and narrative examples.

3.1 Blind moderation with behavioral thresholding

Young people are a major demographic of Internet users and UGC participants. In 2013, 3–17-year olds represented 24.3% of people living with Internet connections (File & Ryan, 2014) and possibly more than a third of cell-phone owners (Duggan & Brenner, 2013). In the US, 25% of 3-year-olds go online daily, and that proportion increases to almost 70% by age 8 (Gutnick, Robb, Takeuchi & Kotler, 2010). Despite the proportional overrepresentation of young people on the Internet, fewer than 20% of studies of Internet use attend to anyone below the age of 9 (Holloway, Green & Livingstone, 2013), and current research falls short of addressing the safety of these users (Hartikainen, Iivari & Kinnula, 2015).

The child-focused use case is interesting from a safety perspective. Strategies that are appropriate for adult populations may not work with children, e.g., Facebook's real-names policy. And yet, it is also not an option to keep children away from the Internet. Only a decade ago, the best practice recommended to parents was to limit young people's exposure to the Internet until they had reached maturity. But the now-current image of "digital natives," a generation who has had access to the Internet since infancy, implies that more direct solutions to the safety of UGC are called for.

The special challenges of designing youth-oriented UGC systems highlight many of the general challenges of online content moderation in general. They also emphasize how moderation and privacy interact: children have peculiar rights to both privacy and safety online. But it would seem impossible to provide one without violating the other; how do you establish the safety of a private letter to its recipient without opening it up and, in that sense, invading the privacy of that recipient?

3.1.1 Definition of behavioral thresholding

We introduce a method for blind, self-moderating, youth-oriented UGC using behavioral thresholds. The central assumption of BT is that any semantic unit that 100% of users would emit must be acceptable to transmit to them all. Though seemingly mundane, this claim is powerful because it allows us to definitively establish the safety of an emission, even knowing nothing of its content. A trivial system for preventing proscribed emissions would only allow the transmission of strings that had been generated and transmitted by 100% of users. But if the claim is true at the threshold of 100% of users, it may be true at 99% or 98%. In such a system, the

"behavioral threshold" for acceptable emissions is this threshold percentage. Behavioral thresholds were introduced by mathematical sociologists to model social dynamics such as opinion change (Granovetter, 1978; Macy, 1991; Schelling, 2006). In practice, it is unlikely that there are any semantic units that are emitted by 100% of users in a text-based UGC system; even a common greeting such as the string "Hi" does not meet this threshold in the chat corpus we analyze in Section 3.1.4. Fortunately, and key to BT, thresholds well below 100% may accomplish the same effect. Even with a very high, say 10%, rate of users who desire to emit a specific bleep, an 11% threshold will be successful in guaranteeing that this problematic content is successfully filtered. But the tests we report below succeed with an even lower threshold, of just 2%.

As we define it, behavioral thresholding is ideal for the moderation of symbol strings, specifically simple text strings in a chat setting. While BT is introduced here for the moderation of chat text, it can work in any UGC platform in which two semantic units emitted by different users can be automatically identified as being the same. Two emoji strings are easier to automatically identify and compare than two freehand drawings of a boat, so UGC in the form of freeform art is unsuited to this method.

Behavioral thresholding may at first glance seem unlikely to be at all effective. In a substantive conversation, we do not make statements that are likely to be exactly repeated by a large proportion of the population. This suggests that the BT method's arbitrarily low false negative, or "miss" rate, should be complemented by a correspondingly high false-positive rate: though being inappropriate should imply that a message is sufficiently rare, being sufficiently rare should not at all imply that a message is inappropriate. In guaranteeing the absence of inappropriate content, BT will incorrectly filter out large numbers of perfectly appropriate statements.

With this higher value on preventing false negatives (true bleeps that incorrectly passed the filter), at the cost of an increased false-positive rate (acceptable emissions that were incorrectly flagged as bleeps), BT suits itself to the task of youth-oriented UGC applications. A young user base is more likely to consist of poor typists with immature language and developing social skills. Developing language users are less likely to draw upon a large lexicon, uncommon syntactic constructions, or other language features that encourage unique statements and undermine thresholding. Their developing social interaction skills are likely to make them less sensitive to the problem of their many below-threshold messages being withheld. More broadly, it is generally accepted that parents have the authority, and perhaps a responsibility, to control and even censor their children's media exposure. The sense of obligation to protect young users from inappropriate content translates to a mandate to minimize false negatives, even at the cost of a correspondingly high rate of false positives. In the youth-oriented use case, the precedence of safety over the other qualities therefore relaxes technical issues endemic to the challenge of automatically moderating UGC.

Narrowing the scope of the design problem to younger people adds some design challenges and relaxes others. Young people are more likely to use non-standard spellings and syntactic constructions. But as long as non-grammatical outputs are customary, BT will be able to aggregate and pass them.

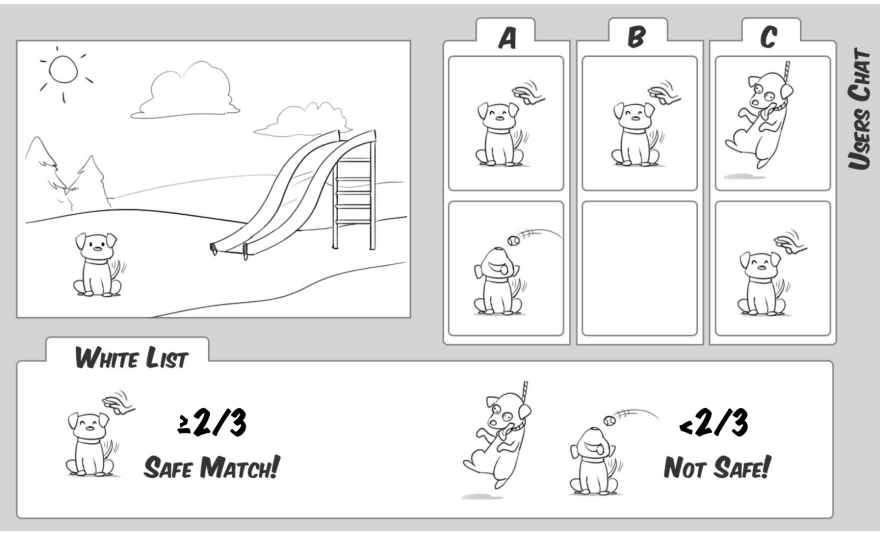

Figure 1. ***Users A, B, and C are interacting with the scene on the upper left by emitting semantic units. User A has emitted "Pet the dog" and "Play fetch." User B has emitted "Pet the dog." User C has emitted "Yank the dog" and "Pet the dog." Only "Pet the dog" exceeds this example's conservative behavioral threshold of 2/3. False positives, such as the omission of "Play fetch," are an inevitable side effect of this method.***

3.1.2 Example 1: Illustrative example

In Figure 1, we have sketched an example of BT in action. Users A, B, and C are interacting by emitting semantic units. They can describe a scene, interact with it, and even elaborate on it by introducing new semantic units. However, only emissions that have exceeded the behavioral threshold are broadcast. This UGC system is expressive because the number of possible semantic units grows combinatorially with the number of basic elements ("dog," "ball," "collar," "hand"). Yet, despite this large UGC space, a unit that has been expressed once remains likely to be expressed again.

Since all three users emitted "pet the dog," that semantic unit can be defined as safe and whitelisted permanently, without moderator approval and, indeed, without any consideration of what it means. The other two semantic units, throwing to the dog and yanking the dog, did not

exceed the threshold and were withheld. Note that one of the filtered emissions is actually benign but too rare to get whitelisted. The conservatism of this method is inherent to it and is a typical and desirable feature of child-focused UGC moderation platforms.

3.1.3 A blind approach to arbitrarily low false-positive rates

For a more formal treatment, take a set of statements X, each with probability $P(bleep)$ of being inappropriate statement "bleep." A bleep can only pass the filter if its frequency meets or exceeds behavioral threshold $T \in [0.00,1.00]$. The probability of this happening is

$$P(\#@\%! \geq T) = \sum_{n=T|X|}^{|X|} P(\#@\%! = n) \tag{1}$$

where the function within the summation is the binomial distributed probability that exactly n statements will have value *bleep*

$$P(\#@\%! = n) = \binom{|X|}{n} P(\#@\%!)^n \left(1 - P(\#@\%!)\right)^{|X|-n} \tag{2}$$

For rare bleeps ($P(bleep) << 1$), this probability decreases exponentially with increases in n, and all but the first terms of the summation are guaranteed to be very small. Since T defines the lower bound of n, increments in T function to remove these initial terms, causing exponential decreases in $P(bleep > T)$. Increasing T by small amounts will quickly make the probability that that bleep escapes the filter negligible. For example, let $|X|=100$ and $P(bleep) = 0.01$. When T is raised slightly from 0.01 through 0.04 to 0.08, then $P(bleep > T)$ drops dramatically from 63%, through 1.8%, to 0.00082%—by roughly an order of magnitude for each percent increase in the threshold. With small adjustments of T, the designer can set the probability of a false negative arbitrarily low. Within this formalism, behavioral thresholding can be seen to generalize the well-known ESP game (von Ahn & Dabbish, 2004; 2008) which implements human-computed image tagging under the special case of a two-person institution with $T=1.00$.

It should be clear that the appealing properties of BT come with many strong assumptions and a very high false-positive rate. The empirical evaluation in Example 2 tests the assumptions and shows how BT can be complemented by more precise methods. But, independent of its role as a potential production system is its role as a proof of the existence of blind moderation.

3.1.4 Example 2: Empirical evaluation and assumption checks on the SafetyText system

BT demonstrates the existence of blind moderation, but, as a proposed system, BT's domain of application is limited. First, it demands a domain such as child-focused UGC in which the cost of a miss is much larger than the cost of a false positive. Second, there must be a corpus of content that is high enough in volume and redundancy that most semantic units have been emitted more than once. Third, bleeps must be rarer than appropriate content. And relatedly, if collusion exists among users who wish to "beat" the system, the size of that group of users must not exceed $I\%$ of the total population of the system (in other words, the existence of malicious colluders impose a practical lower bound on I).

Though these seem like prohibitively strict assumptions—particularly the first two—they hold in a wide range of currently popular child-focused social games and social networks such as Webkinz, Animal Jam, Wizard101, and Fantage, each of which offers "safe chat" systems with artificially constrained lexicons and strict character limits. All of these systems impose constraints on lexicon and syntax that, combined with the simple typing of children, "tab autocomplete" suggestion functionality, and Zipf's law of word frequencies (Zipf, 1949), will send a surprising percentage of messages above threshold. Going beyond these general claims in support of the viability of BT, we can demonstrate it more specifically in the case of a heavily used, proprietary, child-focused UGC moderation system we refer to as SafetyText.

SafetyText is a UGC moderation platform that mediates chat interactions online. It is a popular system that, over a decade, has moderated perhaps 10 billion youth-generated chat messages by hundreds of millions of users, in dozens of countries and multiple languages. SafetyText is based upon an NLP system that uses a restricted syntax and semantically enriched lexicon to implement safety rules that can, for example, prevent one user from derogatorily calling another a "beach." Integrated into the system is a trained moderation staff that provides support by manually moderating samples of chat, integrating with the game culture, refining the lexicon, and arbitrating special cases. Because of the sensitive nature of its target population, the system has a high false-positive rate: it ends up conservatively filtering many messages that are clearly safe in order to more reliably prevent the transmission of the small number of unsafe messages.

The SafetyText NLP system's approach to a message that it cannot identify as definitely inappropriate is to "fake send" it by artificially making it seem to its speaker to have been sent. The fake-send feature might seem to completely undermine the viability of the system, but SafetyText has been at least expressive enough to successfully compete for and maintain a user base of hundreds of millions of young people. Since most users are poor typists whose language and social interaction skills are still developing, it may be that they rarely notice unsent messages, or tolerate them when they do notice.

The robustness of the preexisting SafetyText platform to these weaknesses indirectly supports the assumptions on which the success of the BT method depends. However, we also tested these assumptions directly by evaluating BT on 250,000,000 SafetyText messages, covering a span of multiple months and multiple language groups. We assert that:

- *More than 80% of messages were emitted more than once.*
- *Inappropriate content is indeed rare: thresholds* T *as low as 0.02 are adequate. Among English speakers the first two bleeps, "fuck" and "shit," are seen in 1.5% and 0.5% of users respectively.*
- *SafetyText has a high false-positive rate: 60% of all messages are flagged by the system as conceivably inappropriate and are not sent. This statistic stands despite the objective rarity of inappropriate content, the system's complexity and sophistication, and the system's sustained success and popularity. SafetyText continues to represent the state of the art in youth-oriented UGC moderation.*
  Repeated messages are common enough for BT to be a surprisingly effective heuristic moderation technique. BT technically works at T=*0.02, passing 15% of all chat at negligible computational cost, although this 15% consists of only four unique messages (because all other messages were sent by fewer than 2% of users). Decreasing the threshold to* T=*0.005, the number of messages that exceed the threshold for being sent increases from 4 to about 100 messages. These 100 messages account for 25–40% of the bulk of chat (in the subset of English language chats, BT at this half-percent threshold passes 27.2% of chat), more than ten of which are acceptable "ham" messages that the production SafetyText system rejects as "spam." However, one of those 100 messages is a known bleep, which means either that* T=*0.005 is too low, or BT requires human assistance at this threshold.*

Because BT can automatically vet approximately a third of SafetyText messages, we propose that BT may serve a supporting role within a production system. Even implemented with an excessively low threshold that permits misses, like the T=0.005 threshold above, BT can serve a useful role within a more sophisticated moderation pipeline. Because a small change in the threshold leads to a large change in the proportion of messages passing the system, BT can also be used to flag "high-priority" messages for human inspection. BT may also be adaptable to changes in slang, fads, and other popular language: to accomplish this more challenging task it need only flag novel statements that exceed threshold. Overall, BT could reduce the load on more sensitive and resource-intensive moderation methods by providing a first pass over the bulk of chat. The efficiency of BT is rooted in its complete ignorance of message semantics in favor of attention to social patterns of use.

3.1.5 Example 3: Improved effectiveness if only certain types of players emit bleeps

Though all users are potential emitters of inappropriate content, research shows that some users are much more likely than others to violate community norms (Cheng et al., 2015). Fortunately, BT is particularly effective in environments where norm-violating expressions are peculiar to a specialized "bleeper" type. If a specific bleep is equally likely (say, $P(bleep)$ = 0.01) from each of

100 users, then BT has a 26.4% chance of passing it at T=0.02. But if just 1 of the above 100 users has a 100% probability of emitting bleep—if only bleepers bleep—then that content is guaranteed not to pass, no matter the volume of attempts that single user generates.

3.1.6 Example 4: Moderating encrypted content

Our explication of BT has been in terms of known messages, so in what sense is BT blind? Imagine a system identical to SafetyText except that all messages between pairs of players are end-to-end encrypted with the same public key. This change would confound the efforts of all current schemes for filtering unsafe content, whether based on human moderators or computational NLP schemes. But BT would still be able to provide some indication that certain messages were guaranteed to be safe. Behavioral thresholds are calculated in terms of the frequencies of different semantic units. Because the frequency of a unit can be calculated with only its hash sum, it is possible to implement the method without a system designer ever having access to the actual content of users' private messages. This is an interesting property because it wouldn't seem that content moderation is possible without actual access to content, but BT can issue safety guarantees while respecting privacy, at least in the narrow domain of child-focused UGC.

3.1.7 Limitations

BT is blind, safe, and scalable. It is also expressive in the sense that users may enter any piece of content to the system (even though that content may not survive the filter). But BT has a few weaknesses. Obviously, it is unlikely to be satisfactory as a standalone moderation solution. In practice, even very low thresholds will have false-positive rates that are too high for a sufficiently expressive UGC experience. BT is therefore most likely to find practical use as one component of a more complex, hybridized moderation system, such as SafetyText. Although it is theoretically interesting that BT requires no access to semantics, from a practical perspective it is most likely that a moderation system with access to semantics will aim to leverage that information as well.

BT is restricted to UGC platforms in which a large proportion of content is reproduced exactly by many users. These conditions are unlikely to exist outside of narrow domains, such a child-focused platforms that place artificial bounds on lexicon, syntax, and message length. However, there do exist computational methods for recognizing two non-identical inputs as "the same." For example, locality-sensitive hashing has already proved fruitful in image processing and digital forensics (Bjelland et al., 2014), and it could expand the range of BT beyond UGC platforms based on short text strings.

BT also requires some "burn-in": in practice, a safe statement would be categorized as unsafe until it reached threshold. This problem could also be alleviated in practice by integrating with a preexisting hybrid production system such as SafetyText.

Lastly, the key assumptions of BT restrict it to the somewhat narrow use case of child-focused UGC moderation. Though it seems restrictive, we argue in the discussion that this last property is as much a strength as a weakness, as it prevents BT from being applied abusively in non-youth censorship applications.

3.2 Blind moderation with reverse correlation

Where behavioral thresholding satisfies its assumptions with the requirement that users are children, the second human-computational UGC-moderation method we introduce, reverse correlation, is not for child-only uses, but instead restricts itself to Collective UGC, a type of UGC that is less focused on facilitating personal expression than mass engagement. While the effectiveness of behavioral thresholding depends on the existence of a strict identity relation between emissions by different people, reverse correlation has the advantage of being able to aggregate over contributions with subtle differences.

In Collective UGC, the aggregated efforts of hundreds or even millions of individuals are represented as a single (usually artistic) output. For most Collective UGC, a central planner posts a "central prompt," every participant contributes content, and the product is some aggregation over all contributions. Despite its impressive potential to engage and unite very large populations in the shared production of a common goal, Collective UGC is in practice much less common than other types of UGC, in part because of how current manifestations inefficiently trade off the qualities of expressivity, safety, and scalability. The method we propose circumvents many of these weaknesses.

A prominent example of Collective UGC is the Sheep Market, a 2007 work by Aaron Koblin (Koblin, 2009). Koblin used the Amazon Mechanical Turk crowdsourcing platform to solicit and vet thousands of drawings of sheep. Workers were paid $0.02 to fulfill the central prompt: "draw a sheep facing to the left." The collection of 10,000 solicited drawings of sheep was united with a visual interface that made them easy to browse.

Just as the Sheep Market represents the inspiring potential of Collective UGC projects, it also represents their downsides. Because the communities of collaborators supporting Collective UGC come from a large, uncontrolled, anonymous audience, Collective UGC is vulnerable to vandalism. While contributions can be manually moderated to ensure compliance, this is costly in time or money. In the Sheep Market, human moderation efforts were necessary to ensure that each of the 10,000 drawings was compliant in depicting a left-facing sheep and only that. Whether crowdsourced or not, the task of moderating Sheep Market contributions scales linearly with project size, and the approach is in that sense not scalable.

The Sheep Market may not have been practically or financially viable as anything more than a one-off demonstration. The overhead of moderation would likely have become a problem if it had tried to scale to millions of users, or to continue issuing new prompts on a regular weekly or

monthly basis. But without any moderation, the Sheep Market would have been lower quality and potentially inappropriate for general audiences. A stricter prompt may have been easier to vet automatically (improving scalability) but only at the cost of decreased expressivity. The result of these safety/scalability/expressivity trade-offs is a UGC system that either fails to engage participants, exposes them to un-vetted content, or is not viable at scale. These problems have severely limited the use of Collective UGC beyond prototypes and one-off demonstrations. Fortunately, new human computation schemes can help overcome these weaknesses, and may help increase the viability and popularity of collective UGC as an approach to eliciting large-scale audience engagement.

***Figure 2.** In the reverse-correlation task, a base image (which is itself never exposed) has a filter of random noise added to and also subtracted from it. Users repeatedly choose between the resulting pair of images, in favor of whichever is a better visual representation of some prompt. Figure 3 shows the results for different prompts. See Acknowledgements for all image sources.*

3.2.1 Reverse correlation

We introduce a blind method for aggregating contributions in a Collective UGC setting that is safe, scalable, and expressive. "Reverse correlation" integrates an unlimited number of user-

generated contributions in an expressive medium (such as images or audio) such that the aggregated output is guaranteed to be consistent with the goals of the collective, and specifically, to be free of inappropriate content. Like BT, RC leverages the relative sparsity of inappropriate content to "average out" any off-prompt contributions in favor of the content prescribed by the central prompt. Like BT, RC is blind, a point we develop in Example 3.

RC is already well known in social psychological research on social-reasoning processes (Dotsch & Todorov, 2012; Dotsch, Wigboldus, Langner & van Knippenberg, 2008; Julie, Mangini, Fagot & Biederman, 2006; Karremans, Dotsch & Corneille, 2011; Todorov, Dotsch, Wigboldus & Said, 2011). As with BT, what is novel is the introduction of this already-proven research technique to practical design challenges in human computation.

In an RC task, a template is defined upon which the final output will build (Figure 2). This template can be an image, a sound, or some other representation of an abstract high-dimensional space. Users are never exposed directly to the raw template, but only to two random variations on it. On being presented with the prompt, they select which of two variations on the template is a better representation of the intent of the prompt. Afterward, these random variations are averaged to produce a collective output that, with more inputs, increasingly resembles the output specified by the prompt. While most Collective UGC applications follow the Sheep Market in aggregating all inputs into a large collection of independently viewable contributions (e.g., a grid of adjacent images), RC aggregates those inputs in a large collection of juxtaposed pixels that constitute a single collective contribution (e.g., a stack of image layers).

3.2.2 Example 1: Eliciting diverse mental images

In the example in Figure 2, a "filter" of random noise has been added to and subtracted from a single template image of a face, producing two variants that are presented to the user as options. The user is then solicited with the central prompt, which may read: "Please select the face that looks more happy," or "more scary," or "more like a Moroccan person." Neither variation will look at all happy or scary or Moroccan, since each is only a noisy version of the template. However, if the prompt is repeated hundreds of times, the average of all selected variants will resemble the face specified in the central prompt (Figure 3). The method is sensitive enough to reflect noticeable differences in peoples' "mental images." For example, researchers have found that, when racially prejudiced subjects are asked to produce out-group faces, outside raters judge the resulting faces as looking less trustworthy and more criminal (Dotsch et al., 2008).

RC may be seen as "search" through a high-dimensional UGC space. Dotsch's (Dotsch & Todorov, 2012) images used a base face as a template, and their subjects' choices collectively effected search through the 4092-dimensional space of visual noise parameters. Despite this very high dimensionality, researchers have demonstrated satisfactory results with "search" over only hundreds of choices, making reverse correlation practical to implement with or without a very large-scale participant pool.

With proper implementation, choices that violate the binary intentions of the central prompt will be rare and incoherent, and will therefore be averaged out to form a neutral palimpsest background against which a final output can emerge that is consistent with the prompt. Our argument here follows the same tack as that formalized for BT: if the probability of a single violation is sufficiently small, then the probability that independent agents will commit precisely the same violation is negligible, and through thresholding (in the case of BT) or averaging (in the case of RC) the system will perform moderation automatically.

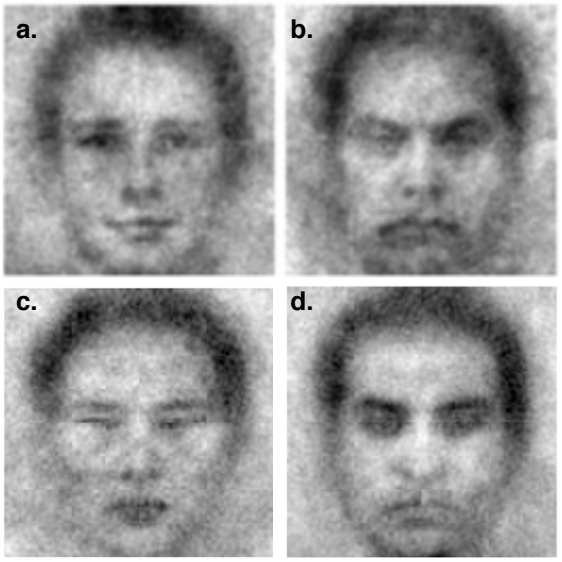

Figure 3. ***Typical outputs of the reverse-correlation task, for prompt: "Which of these two faces looks more _____." Prompts included a. trustworthy, b. untrustworthy, c. Chinese, and d. Moroccan. Each of these outputs is the result of only a few hundred binary choices.***

3.2.3 Example 2: Audio RC with a physical interface

The binary decision setting required by RC is simple and flexible enough that the method is easily adapted to many problem domains. In fact, RC was originally developed not for images, but for the generation of evocative audio clips. The method thus provides a simple design pattern for integrating UGC into the design of a larger multimedia user experience. Assume, for example, that a designer wants to leave the domain of the Internet and foster a sense of ownership in visitors to a physical haunted castle amusement-park attraction. Using RC, castle visitors might encounter two mysterious doors at the end of a corridor, and be given the option to go either left or right (Figure 4). Each door would have an automatically generated face (or a sound) posted

above it, and the only difference between the doors would be the random difference between these faces. If visitors were instructed to avoid the more frightening of the two faces, then their behavior would constitute a choice in favor of the avoided filter, which would then be used to make subsequent versions of the face more frightening for future guests. Guests can derive value from the knowledge that the scary face at the haunted castle is in some sense customized by their own frightening experience there.

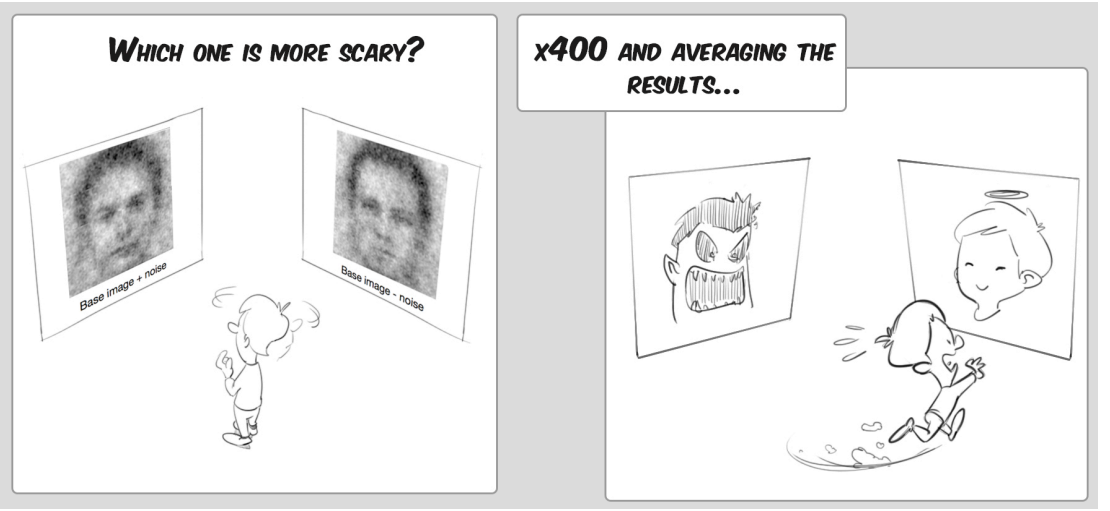

Figure 4. ***Representation of the reverse-correlation task in a haunted castle scenario. The average of a few hundred binary reactions to the prompt results in a stimulus that is increasingly consistent with that prompt. Because each decision can be made by a different person, this method can be used for the development of Collective UGC. The aggregation implements safety without any human moderation.***

3.2.4 Example 3: Deposing the tyrant

It should be clear that RC is safe, scalable, and expressive, but in what sense is it blind? Permitting a UGC system to average all submitted images implies access to those images. RC can be parallelized and distributed over a cluster, will the *n*th pixel of each image being averaged on core *n*, but such an effort is a poor implementation of privacy if the moderator has an owner's access to the cluster. However, there remains a weaker but still legitimate sense in which RC is blind. RC requires access to raw content, but not to content semantics.

Imagine an election between an incumbent dictator and a democratic favorite. All citizens prefer the latter but fear their ballots could be traced. If RC was used for the voting mechanism, with

voters selecting the one of two noisy faces that most closely resembled their preferred candidate, then they could vote in safety: even if the tyrant gained access to all choices, and even if the collective average resembles the challenger, the noisiness of the binary choice setting would make it impossible to reliably infer any one voter's preferred candidate from their chosen image. Because RC's two-alternative forced choice is between options with individually meaningless differences, RC obscures users' intentions, and thereby protects their privacy. So while RC is less blind than BT in the sense of forgoing access to a message, it is still blind in the sense that it functions without access to a message's meaning.

As a proposed system, RC is likely a poor improvement on existing voter protection schemes, which have long been interested in the preservation of privacy and anonymity (Chang & Lee, 2006; Fujioka, Okamoto & Ohta, 2005). Furthermore, the example of the tyrant has no sense of inappropriate content (write-in candidates?), but this example highlights how a blind human-computational approach can aggregate collective intentions while still obscuring individual ones, and how such an approach might differ from current computer scientific voting schemes for preserving privacy.

3.2.5 Limitations

RC shares with BT the weakness that it is vulnerable to large-scale collusion: if conspirators constitute a large enough proportion of users, they may be able to inject off-prompt content into the collective output. Fortunately, this level of collusion presents vandals with a substantial collective active problem. Additionally, Collective UGC is at present only an exotic variety of UGC, making the scope of applicability of RC narrow. That practical problem does not, however, detract from RC as a second proof of the existence of blind UGC moderation in certain constrained problem domains. With reverse correlation for collective UGC, any design of these descriptions can be used to safely, expressively, scalably, and blindly integrate interactivity into online content.

4. DISCUSSION

The behavioral-threshold method uses the relatively low frequency of inappropriate emissions and of people who emit them to offer formal guarantees that an inappropriate message will not be visible to users. Though this method may be applicable to youth-focused platforms, it is less likely to be effective within a platform that serves a general audience. This is because there are differences in what different user communities tolerate. For audiences of children, misses are very costly, and false positives are tolerated at surprisingly high rates, as in the case of a popular system we evaluate above, that we estimate inappropriately withholds twelve safe messages for every unsafe one it successfully catches. BT is a fit for this specific use case because it allows an arbitrarily low miss rate, at the cost of the kind of high false positive rate that young audiences seem to tolerate. The freedom to focus only on reducing the false negative rate, with little regard

to corresponding increases in the false positives, allows us to consider designs that would not work in adult domains

Using the reverse-correlation approach to Collective UGC, a group of users can effectively search an unlimited space of possible images in a way that protects their members from exposure to content that diverges from a predefined prompt. Like BT, it relies on the assumption that inappropriate emissions are rare, but it works from this foundation in a different way. RC computes a single output by averaging user inputs, all of which are elicited in response to the same prompt. Assuming that most users' inputs are in line with the prompt, or that any deviations from the prompt are statistically unrelated to one another, averaging the inputs creates a coherent on-prompt output (such as an image) that automatically smooths off-prompt contributions into the background.

4.1 Comparing and contrasting behavioral thresholding and reverse correlation

BT and RC share a number of appealing properties. Both rely on the uncommonness of bleeps to exclude them from the output; both permit expressive UGC in diverse large-scale media without human moderation load; neither depends on knowledge of content semantics; and both can be integrated into the more complex pipelines characteristic of real-world moderation systems.

Despite their similarities, BT and RC are, of course, different methods with different use cases. BT leverages the rarity of bleeps to excise them, while RC leverages this rarity to average bleeps out. BT is intended for youth-oriented cases that tolerate a high false-positive rate, while RC is suited only to the subset of UGC systems focused on Collective UGC, in which many users' inputs are aggregated into a single collectively produced output.

The methods also differ a bit in how they provide privacy and expressivity. BT can respect the privacy of users in the straightforward sense that it can accept hashes that represent content in place of raw content. RC respects the privacy of users in the different sense that an observer viewing a single user's contribution cannot reliably infer that user's position with respect to the prompt. Also, while both methods are in some sense "expressive," they differ in what sense. In BT, users can submit any message content into the system, but little of that content will be passed through it. Conversely, in RC, users are restricted in the content that they can submit into the system, but it places no constraints on the output space. Within this contrast between the two methods, BT might be described as "upstream expressive" and not "downstream expressive," while RC is "downstream expressive" but not "upstream expressive."

Looking more closely into the similarities between the methods highlights our main contributions. Overall, each method leverages the unique properties of their domain—child-focused and collective formats for UGC—to blindly implement content moderation. Like many techniques in human computation, these methods circumvent the complexities of natural language

processing by relying on behavioral rather than semantic structures. In the process, they introduce problem settings in which it is possible to provide message monitoring while respecting message privacy.

4.2 Limitations

Aggregating the limitations reported in sections 3.1.7 and 3.2.5, BT and RC have a few weaknesses and vulnerabilities in common. Both are defined in terms of constrained UGC use cases; both are unlikely to be satisfactory as 100% standalone moderation solutions; and both may be vulnerable to coordinated efforts to undermine their effectiveness. Specifically, since both methods assume the statistical independence of individual contributions, they are both susceptible to workarounds such as large-scale collusion. Neither can successfully filter bleeps within user communities that might organize, intentionally or unintentionally, to embrace and emit popular bleeps widely. In the case of reverse correlation, even a single individual with many opportunities to contribute choices could possibly collude with him or herself to pollute the final product with off-prompt content. Fortunately, this level of collusion presents aspiring vandals with a substantial collective active problem.

4.3 BT and RC in social context

While this work is focused on the theoretical possibility of blind UGC moderation, it is worth exploring how these methods might function in the real world. In application, both are most likely to be used in combination with other approaches, in particular those that involve professional human moderators. Both methods are likely to become more effective, in terms of both accuracy and cost, when they are used to reduce but not remove the need for occasional acts of costly manual moderation by humans. Human moderators provide a sanity check on these systems, which are probably vulnerable to exploits based upon large-scale collusion. With complementary human staff, a production system could more robustly reduce both the false-positive and false-negative rates to which BT or RC alone might be susceptible.

The sensitivity of our methods to social context permits us to leave definitions of propriety to the communities managing these systems. But in doing so, these methods may put more importance on the ability of communities to successfully govern their own affairs. If some item of content is appropriate within a community, but a bleep outside of it, then part of limiting the exposure of members to content that they experience as inappropriate will be the development of boundaries, such as membership procedures, or acculturation processes, such as onboarding procedures, that preserve the in-group's culture and its specific sense of what constitutes appropriate content (Kraut & Resnick, 2012, ch. 5).

Considering broader social context also raises ethical questions. For example, there may be no clear line between a content-moderation system and an overt censorship system. Researchers in the domain of UGC moderation have an ethical responsibility to consider how their methods can

be abused by "enemy of the Internet" states and other autocratic authorities (Hartley, Lumby & Green, 2009). Indeed, Akdeniz reports on the frequency with which web-filtering systems are applied beyond their publicly reported scope to implement censorship illegally or unethically (Akdeniz, 1998). Both practical and ethical concerns demand an intentional design approach that remains sensitive to the potential for abuse, and to the unique characteristics of a target population (von Ahn & Dabbish, 2008; Flanagan, Howe & Nissenbaum, 2005).

RC and BT show that integrating the unique properties of a specific population into the design of a moderation system can prevent that system from being abusively deployed on other populations. Both transfer poorly outside of their narrow use cases. Consider the potential for abuse in state-sponsored censorship schemes. The very high false-positive rate that makes BT useful in youth-oriented domains makes it poorly suited to more general applications. The young populations over which BT does implement "censorship" are populations for which it seems acceptable, or even necessary, to do so; parents are often expected to censor their children's media exposure. In the case of RC, collective UGC is a relatively obscure type of UGC setting whose intersections with state-sponsored censorship, if they exist at all, are not clear.

A second ethical concern is the "human cost" of professional content moderation, by which paid moderators suffer from exposure to distressing content (Chen, 2014). Crowdsourcing the work of these professionals may only increase the variety of negative consequences experienced by workers (Silberman, Irani & Ross, 2010). In our methods, content producers are implicitly their own moderators: in behavioral-thresholding systems, the only individuals who will ever be exposed to inappropriate content will be those who produced it, while in reverse-correlation systems, not even the individual who attempts to subvert the content stream is likely to actually observe off-prompt content.

4.4 Future directions

The goal of this work was less to present production-ready moderation systems than to demonstrate the existence of methods that, within their narrow domains, accomplish the seemingly impossible task of blind moderation. Nonetheless, it is useful to consider how these methods might be improved. The false-positive rate of BT quickly becomes untenable outside of the heavily constrained chat-based UGC systems characteristic of youth-focused domains, while most previous work demonstrating the effectiveness of RC has restricted itself to a single type of image stimulus, namely faces. Future work could extend the viability of both methods to a wider class of methods, or identify problem domains whose constraints make other human-computational approaches to blind UGC moderation possible.

More broadly, the conceptual framework on which this work relies remains somewhat underdeveloped. While we discussed some of the interactions between features such as safety, privacy, expressivity, and scalability, our discussion was almost certainly incomplete, and future work should more rigorously distinguish between conditions in which these features must be

traded off and conditions in which improvements in one dimension do not come at the expense of effectiveness in another.

More importantly, there are likely other dimensions along which UGC systems vary. One class of UGC moderation systems that we did not discuss at all introduces variation along a dimension that we might call "peer governance." This dimension captures whether acts of moderation are performed by professional staff, on one extreme, or completely by user peers on the other. It is increasingly popular to empower users to govern content themselves. Any site that lets users report each other's bleeps is implementing peer governance to some degree. Systems that rely more heavily on peer governance include Wikipedia, Reddit, and sites in the Stack Overflow network. There are even reports, on child-focused UGC platforms such as the online social game Webkinz, of successful self-governance among communities of children and tweens (Kafai & Searle, 2010). That said, there are to date vastly more failures of self-governing UGC systems than successes. The complexity of designing community-governance systems, especially of designing systems that design themselves by empowering users to vote upon governance rules, makes these approaches difficult to replicate and unpredictable to manage. User communities can gain surprising political leverage over a platform when they are placed in essential governance roles, as in a recent mass strike of community moderators on Reddit, one that caused dramatic drops in traffic to the site and, ultimately, the resignation of the company's CEO (Tox77, 2015). While exciting advances are being made in the study of community-governed UGC (Frischmann, Madison & Strandburg, 2014; Heaberlin & DeDeo, 2016; Kraut & Resnick, 2012; Lampe, Zube, Lee, Park & Johnston, 2014; Mueller et al., 2015; Müller et al., 2015; Schweik & English, 2012), this style of solution is still poorly understood and, therefore, not easily replicable from platform to platform, nor easy to integrate into the framework that we use to motivate the BT and RC methods.

5. CONCLUSION

Though it may seem futile to try moderating a piece of content without being able to analyze it, social patterns make such "blind" moderation possible. Drawing from mathematical sociology and computational social psychology, we introduce two methods to human computation, behavioral thresholding and reverse correlation, both of which can offer UGC moderation that is not only blind, but safe, expressive, and scalable. This work demonstrates the power of human computation, social psychology, and other behaviorally focused methods as sources of insight for designing human computation institutions.

6. ACKNOWLEDGMENTS

The authors wish to thank the anonymous reviewers for their helpful comments. Figures 2 and 3 were adapted from figures published in (Dotsch & Todorov, 2012; Todorov et al., 2011), reprinted with permission from Ron Dotsch.

7. REFERENCES

Adler, B. T. & de Alfaro, L. (2007). A content-driven reputation system for the Wikipedia (p. 261). Presented at the ACM WWW, New York, New York, USA: ACM Press. http://doi.org/10.1145/1242572.1242608

Ahn, von, L. & Dabbish, L. (2004). Labeling images with a computer game (pp. 319–26). Presented at the ACM EC, New York, New York, USA: ACM Press. http://doi.org/10.1145/985692.985733

Ahn, von, L. & Dabbish, L. (2008). Designing games with a purpose. Communications of the ACM, 51(8), 57–67. http://doi.org/10.1145/1378704.1378719

Akdeniz, Y. (1998). Who watches the watchmen? Internet content rating systems and privatised censorship. The Australian Library Journal, 47(1), 28–42. http://doi.org/10.1080/00049670.1998.10755831

Barnes, S. B. (2006). A privacy paradox: Social networking in the United States. First Monday, 11(9). http://doi.org/10.5210/fm.v11i9.1394

Belenkiy, M., Chase, M., Erway, C. C., Jannotti, J., Küpçü, A., Lysyanskaya, A. & Rachlin, E. (2007). Making p2p accountable without losing privacy (p. 31). Presented at the 2007 ACM workshop, New York, New York, USA: ACM Press. http://doi.org/10.1145/1314333.1314339

Bjelland, P. C., Franke, K. & Årnes, A. (2014). Practical use of Approximate Hash Based Matching in digital investigations. Digital Investigation, 11, S18–S26. http://doi.org/10.1016/j.diin.2014.03.003

Burmester, M., Desmedt, Y., Wright, R. N. & Yasinsac, A. (2006). Accountable Privacy. In C. B, C. B, M. J. A & R. M (Eds.), Lecture Notes in Computer Science (Vol. 3957, pp. 83–95). Springer, Berlin. http://doi.org/10.1007/11861386_10

Chang, C.-C. & Lee, J.-S. (2006). An anonymous voting mechanism based on the key exchange protocol. Computers & Security, 25(4), 307–314. http://doi.org/10.1016/j.cose.2006.02.004

Chen, A. (2014, October). The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed. Retrieved August 2015, from http://www.wired.com/2014/10/content-moderation/

Cheng, J., Danescu-Niculescu-Mizil, C. & Leskovec, J. (2015). Antisocial Behavior in Online Discussion Communities. Ninth International AAAI Conference on Web and Social Media. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10469

Das, R. & Vukovic, M. (2011). Emerging theories and models of human computation systems (pp. 1–4). Presented at the 2nd international workshop, New York, New York, USA: ACM Press. http://doi.org/10.1145/2030100.2030102

Davidson, J. (2015, July). Facebook's Zuckerberg Defends Controversial "Real Name" Policy. Retrieved September 2015, from http://time.com/money/3942997/facebook-real-name-policy/

Dotsch, R. & Todorov, A. (2012). Reverse Correlating Social Face Perception. Social Psychological and Personality Science, 3(5), 562–71. http://doi.org/10.1177/1948550611430272

Dotsch, R., Wigboldus, D. H. J., Langner, O. & van Knippenberg, A. (2008). Ethnic Out-Group Faces Are Biased in the Prejudiced Mind. Psychological Science, 19(10), 978–80. http://doi.org/10.1111/j.1467-9280.2008.02186.x

Duggan, M. & Brenner, J. (2013). The Demographics of Social Media Users — 2012. Pew Research Center.

Ekstrom, M. & Ostman, J. (2015). Information, Interaction, and Creative Production: The Effects of Three Forms of Internet Use on Youth Democratic Engagement. Communication Research, 42(6), 796–818. http://doi.org/10.1177/0093650213476295

File, T. & Ryan, C. (2014). Computer and Internet Use in the United States: 2013 (No. ACS-28). census.gov. U.S. Census Bureau.

Flanagan, M., Howe, D. C. & Nissenbaum, H. (2005). Values at play (pp. 751–60). Presented at the ACM SIGCHI, New York, New York, USA: ACM Press. http://doi.org/10.1145/1054972.1055076

Frischmann, B. M., Madison, M. J. & Strandburg, K. J. (Eds.). (2014). Governing Knowledge Commons. Oxford University Press.

Fujioka, A., Okamoto, T. & Ohta, K. (2005). A practical secret voting scheme for large scale elections (Vol. 718, pp. 244–51). Lecture Notes in Computer Science.

Gates, C. E. (2007). Access Control Requirements for Web 2.0 Security and Privacy. Presented at the Proc. of Workshop on Web 2.0 Security & Privacy.

Ghosh, A. & McAfee, P. (2011). Incentivizing high-quality user-generated content (pp. 137–46). Presented at the ACM WWW, New York, New York. http://doi.org/10.1145/1963405.1963428

Ghosh, A., Kale, S. & McAfee, P. (2011). Who moderates the moderators? (p. 167). Presented at the ACM EC, New York, New York, USA: ACM Press. http://doi.org/10.1145/1993574.1993599

Granovetter, M. (1978). Threshold models of collective behavior. American Journal of Sociology, 1420–43.

Grimmelmann, J. (2015). The Virtues of Moderation, 17, 42–109.

Gutnick, A. L., Robb, M., Takeuchi, L. & Kotler, J. (2010). Always connected: the new digital media habits of young children. New York: The Joan Ganz Cooney Center at Sesame Workshop.

Harrison, J. (2010). User-generated content and gatekeeping at the BBC hub. Journalism Studies, 11(2), 243–56. http://doi.org/10.1080/14616700903290593

Hartikainen, H., Iivari, N. & Kinnula, M. (2015). Children and Web 2.0: What They Do, What We Fear, and What Is Done to Make Them Safe. In Lecture Notes in Business Information Processing (Vol. 223, pp. 30–43). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-21783-3_3

Hartley, J., Lumby, C. & Green, L. (2009). Untangling the Net: The Scope of Content Caught By Mandatory Internet Filtering (No. 39549). Internet Industry Association.

Heaberlin, B. & DeDeo, S. (2016). The Evolution of Wikipedia's Norm Network. Future Internet, 8(2), 14. http://doi.org/10.3390/fi8020014

Hermida, A. & Thurman, N. (2007). Comments please: How the British news media are struggling with user-generated content. Presented at the 8th International Symposium on Online Journalism.

Hidalgo, J. M. G., Sanz, E. P., García, F. C. & De Buenaga Rodríguez, M. (2009). Web Content Filtering. In Advances in Computers: Social Networking and the Web (Vol. 76, pp. 257–306). Elsevier. http://doi.org/10.1016/S0065-2458(09)01007-9

Holloway, D., Green, L. & Livingstone, S. (2013). Zero to eight. Young children and their Internet use. eprints.lse.ac.uk. LSE, London: EU Kids Online.

Hughey, M. W. & Daniels, J. (2013). Racist comments at online news sites: a methodological dilemma for discourse analysis. Media, Culture & Society, 35(3), 332–47. http://doi.org/10.1177/0163443712472089

Jain, S. & Parkes, D. C. (2009). The role of game theory in human computation systems (p. 58). Presented at the ACM SIGKDD Workshop, New York, New York, USA: ACM Press. http://doi.org/10.1145/1600150.1600171

Julie, M.-M., Mangini, M. C., Fagot, J. & Biederman, I. (2006). Do humans and baboons use the same information

when categorizing human and baboon faces? Psychological Science, 17(7), 599–607.

Kafai, Y. B. & Searle, K. A. (2010). Safeguarding Play in Virtual Worlds: Designs and Perspectives on Tween Player Participation in Community Management. International Journal of Learning and Media, 2(4), 31–42. http://doi.org/10.1037/0012-1649.42.3.395

Karremans, J. C., Dotsch, R. & Corneille, O. (2011). Romantic relationship status biases memory of faces of attractive opposite-sex others: Evidence from a reverse-correlation paradigm. Cognition, 121(3), 422–26. http://doi.org/10.1016/j.cognition.2011.07.008

Kelly, K. (2009, March). Overheard@GDC09: TTP = Time To Penis. Retrieved September 2015, from http://www.engadget.com/2009/03/24/overheard-gdc09-ttp-time-to-penis/

Koblin, A. M. (2009). The sheep market (pp. 451–52). Presented at the Proceeding of the seventh ACM conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/1640233.1640348

Kollock, P. & Smith, M. (1996). Managing the virtual commons. Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives, 109–128.

Kraut, R. E. & Resnick, P. (2012). Building Successful Online Communities: Evidence-Based Social Design. MIT Press.

Lampe, C., Zube, P., Lee, J., Park, C. H. & Johnston, E. (2014). Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. Government Information Quarterly, 31(2), 317–26. http://doi.org/10.1016/j.giq.2013.11.005

Levin, A. & Abril, P. S. (2008). Two notions of privacy online. Vand. J. Ent. & Tech. L., 11, 1001.

Macy, M. (1991). Chains of cooperation: Threshold effects in collective action. American Sociological Review, 56(6), 730–47.

McIntyre, T. J. (2012). Child Abuse Images and Cleanfeeds: Assessing Internet Blocking Systems. In I. Brown (Ed.), Research Handbook On Governance Of The Internet. Elsevier.

McWhertor, M. (2008, June). When Spore Penis Monsters Attack. Retrieved May 24, 2016, from http://kotaku.com/5017350/when-spore-penis-monsters-attack

Mont, M. C., Pearson, S. & Bramhall, P. (2003). Towards Accountable Management of Identity and Privacy: Sticky Policies and Enforceable Tracing Services (p. 377). Presented at the DEXA '03, IEEE Computer Society.

Mueller, S., Solenthaler, B., Kapadia, M., Frey, S., Klingler, S., Mann, R., et al. (2015). HeapCraft: Interactive Data Exploration and Visualization Tools for Understanding and Influencing Player Behavior in Minecraft. Presented at the ACM MIG.

Müller, Frey, Kapadia, Klingler, Mann, Solenthaler, et al. (2015). Quantifying and Predicting Collaboration in Shared Virtual Worlds. Presented at the Proceedings of Artificial Intelligence and Interactive Digital Entertainment.

Nissenbaum, H. (2004). Privacy as Contextual Integrity. Washington Law Review, 79.

Nissenbaum, H. (2009). Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford University Press.

Ostrom, E. (2005). Understanding institutional diversity. Princeton University Press.

Pang, J. & Zhang, Y. (2015). A new access control scheme for Facebook-style social networks. Computers & Security, 54(C), 44–59. http://doi.org/10.1016/j.cose.2015.04.013

Pantola, A. V., Pancho-Festin, S. & Salvador, F. (2011). TULUNGAN: A Consensus-Independent Reputation System for Collaborative Web Filtering Systems. Science Diliman, 23(2), 17–39.

Pearson, S., Rao, P., Sander, T., Parry, A., Paull, A., Patruni, S., et al. (2009). Scalable, accountable privacy

management for large organizations (pp. 168–75). Presented at the 13th EDOCW, IEEE. http://doi.org/10.1109/EDOCW.2009.5331996

Purslow, M. (2015, May). LEGO Universe couldn't deal with the cost of the penis police. Retrieved May 2015, from http://www.pcgamesn.com/lego-universe-couldn-t-deal-with-the-cost-of-the-penis-police

Sayaf, R. & Clarke, D. (2013). Access Control Models for Online Social Networks. IGI Global.

Schelling, T. C. (2006). Micromotives and macrobehavior. W. W. Norton & Company.

Schweik, C. M. & English, R. C. (2012). Internet Success: A Study of Open-Source Software Commons. MIT Press.

Shehab, M., Squicciarini, A., Ahn, G.-J. & Kokkinou, I. (2012). Access control for online social networks third party applications. Computers & Security, 31(8), 897–911. http://doi.org/10.1016/j.cose.2012.07.008

Silberman, M. S., Irani, L. & Ross, J. (2010). Ethics and tactics of professional crowdwork. XRDS: Crossroads, the ACM Magazine for Students, 17(2), 39. http://doi.org/10.1145/1869086.1869100

Silverstein, J., Nissenbaum, H., Flanagan, M. & Freier, N. G. (2006). Ethics and children's information systems. Proceedings of the Association for Information Science and Technology, 43(1), 1–7. http://doi.org/10.1002/meet.14504301133

Sood, S., Antin, J. & Churchill, E. (2012). Profanity use in online communities (p. 1481). Presented at the ACM EC, New York, New York, USA: ACM Press. http://doi.org/10.1145/2207676.2208610

Squicciarini, A. C., Shehab, M. & Paci, F. (2009). Collective privacy management in social networks (pp. 521–530). Presented at the 18th international conference, New York, New York, USA: ACM Press. http://doi.org/10.1145/1526709.1526780

Squicciarini, A. C., Shehab, M. & Wede, J. (2010). Privacy policies for shared content in social network sites. The VLDB Journal, 19(6), 777–96. http://doi.org/10.1007/s00778-010-0193-7

Stefanovitch, N., Alshamsi, A., Cebrian, M. & Rahwan, I. (2014). Error and attack tolerance of collective problem solving: The DARPA Shredder Challenge. EPJ Data Science, 3(1), 13–27. http://doi.org/10.1140/epjds/s13688-014-0013-1

Subramaniam, M., Valdivia, C., Pellicone, A. & Neigh, Z. (2014). Teach Me and Trust Me: Creating an Empowered Online Community of Tweens and Parents. In Maxi Kindling & Elke Greifeneder (Eds.). Presented at the iConference 2014 Proceedings: Breaking Down Walls. Culture - Context - Computing, iSchools. http://doi.org/10.9776/14078

Todorov, A., Dotsch, R., Wigboldus, D. H. J. & Said, C. P. (2011). Data-driven Methods for Modeling Social Perception. Social and Personality Psychology Compass, 5(10), 775–91. http://doi.org/10.1111/j.1751-9004.2011.00389.x

Tox77. (2015, July). Ellen Pao Resigns as Reddit Interim CEO After User Revolt: technology. Retrieved 9/2015, from https://www.reddit.com/r/technology/comments/3cud9k/ellen_pao_resigns_as_reddit_interim_ceo_after/

van den Berg, B., Pötzsch, S., Leenes, R., Borcea-Pfitzmann, K. & Beato, F. (2011). Privacy in Social Software. (J. Camenisch, S. Fischer-Hübner & K. Rannenberg, Eds.) (pp. 33–60). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-20317-6_2

Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H. & Zhao, B. Y. (2012). Social Turing Tests: Crowdsourcing Sybil Detection. arXiv.org.

Yuen, M.-C., Chen, L.-J. & King, I. (2009). A Survey of Human Computation Systems (pp. 723–28). Presented at the International Conference on Computational Science and Engineering, IEEE. http://doi.org/10.1109/CSE.2009.395

Zipf, G. K. (1949). Human Behavior and The Principle of Least Effort: An Introduction to Human Ecology. Reading, Mass., Addision-Wesley.