Walking the Line in Computational Creativity:

Paradox and Pitfall in the Mitigation of Automated Offense

Tony Veale

School of Computer Science, University College Dublin, Ireland.

Tony.Veale@UCD.ie

Abstract

Social media is now the megaphone of choice for many digital-age provocateurs. Social networks offer a wealth of examples of egregious misbehaviour by humans, but perhaps more worryingly, they also serve to magnify and weaponize the automated outputs of generative systems. So what responsibilities do the creators of these systems bear for the offenses caused by their creations, and what can they do to prevent, or mitigate, the worst excesses? For those who "make things that make things," the issue is further complicated by the potential of our systems to accidentally tread on the feelings of others. When does a norm become a troubling stereotype, and how should a system built on social norms avoid crossing the line? We argue that because this line is so easy to cross, our systems should aim to engage with and mitigate offense rather than try vainly to prevent it completely.

Clockwork Lemons

If beauty is in the eye of the beholder, then so, frequently, is offense. Although a shared culture imposes many common values and norms, each of us brings our own sensitivities to bear when judging aspects of creativity and matters of taste. Polite society lays down some bold red lines, but we draw our own lines too, around the acceptable use of words and symbols, around the meanings that those signs can convey, and around the unspoken assumptions that underpin them. As social networking tools extend the reach of our digital provocations, they increase the chances that others will see them – and as a consequence *us* – as objectionable. The same potential for offense increasingly attaches to the outputs of our generative systems, especially when those outputs use much the same signs to evoke much the same meanings. So in what ways are these systems beholden to their beholders, and where do we draw the line so as to minimize offense?

To lean on a cinematic cliché, the outputs of a generative system run the gamut from the good to the bad to the ugly. The "good" represents the idealized targets of our system, a body of outputs exhibiting both novelty and usefulness that embody the desirable qualities of an *inspiring set* (Ritchie, 2007). The "bad" represents cases in which the system falls short of those ideals, with outputs that cleave too closely to the landmarks of an inspiring set – imitation and pastiche – or land too far away to be valued as meaningful or useful. While the "ugly" represents the worst of the rest, the set of outputs most likely to cross the line and cause real offense, it also overlaps as much with the "good" as with the "bad." Ugliness is not the absence of beauty, nor is it its opposite. These are complementary qualities that can sit side by side in the same outputs. One can craft an elegant turn of phrase or a masterly image to express a truly abhorrent viewpoint. If the *ugly* were just a subset of the *bad*, we could reduce the potential for offense by simply making a system better at what it is meant to do. However, as one can sit cheek by jowl with the other, an appealing surface is no guarantee that an appalling message is not lurking beneath. Ventura and Gates (2018) propose that, for certain kinds of creative system – such as those that generate policies and plans – ethical issues can be quantified much like aesthetic issues, with objective functions that score different kinds of value. In a semiotic system of cultural signifiers, however, ethical issues are often subtle, subjective, and difficult to quantify.

At the most general level, there are two broad approaches to regulating offense: *inner* regulation and *outer* regulation. Inner regulation of a generative system curtails a state space and the rules for exploring it, so as to prevent offenses from being committed in the first place. Such a regulator aims to remove offensive capability from a system at the outset, so that it can never even imagine an ugly possibility, perhaps because it lacks the words, concepts or other tools to do so. Outer regulation sees this hope as naive, since most offense is emergent: it is not a property of the parts but of the whole. An outer regulator is an in-built censor that allows a system to generate what it will, before making ethical judgements about which outputs are safe enough to see the light of day. In Freudian terms, inner regulators curb the *id* of a system, while outer regulators impose a *super-ego* to filter this id.

To navigate these options, the paper is laid out as follows. We start by revisiting some extreme examples of offensive online content and/or behaviour, by humans *and* machines. Where did these creators go wrong, and what could an outer regulator have done to prevent them? We then consider the smallest possible ways in which our systems can offend, by surveying some of the simplest generative systems, or *bots*. Bots allow us to unpick content and behaviour, to see that it is often a combination of the two that causes offense. Naïve regulators focus on the parts, not on the whole, as shown by our survey of dictionary-based approaches. Only by shifting to a more holistic viewpoint can we appreciate how offense is all too easily generated by systems with no obvious bias or malicious intent. We must thus recognize the potential for a system's innocent choices to be interpreted as deliberate provocations, since audiences often place a more nuanced spin on outputs than a system itself knows how to generate.

There are many reasons why we in the community build systems whose generative reach exceeds their interpretative grasp. We do so to make the most of meagre resources, or to allow for occasions when a system surprises us much as its target audience. We do so when a system is rich in data but poor in knowledge, or just because shallow generation is easier than deep interpretation (a clever framing strategy can convince an audience of the presence of the latter; see Charnley *et al*., 2012). We do so when the divide between a system's reach and its grasp impresses much more than it disappoints, as when it fosters more *Eliza* than *Tale-Spin* effects (Wardrip-Fruin, 2009). The *Tale-Spin* effect, named after the seminal story generation system (Meehan, 1977) shows a system mashing its gears and producing the wrong kind of sparks, while the *Eliza* effect, named for the famous (and infamous) chatbot, allows systems to take credit where none is deserved. As shown by Weizenbaum's *Doctor* script for *Eliza* (1966), users often ascribe insight and intent where there is none. Yet this effect cuts both ways. If designers are willing to accept *good* Eliza effects, they must also accept its *ugly* effects too, as when insight and intent are ascribed to a system's accidentally insensitive or offensive outputs.

No solution to offense avoidance is without pitfalls, and most breed strange paradoxes of their own. For instance, an outer regulator of vulgarity must also be an inner vulgarian, so as to detect and filter what should not be said. Likewise, a regulator of racist or sexist tropes must be well-versed in the ways of belittling others, so as to avoid the same traps. A regulator aiming to prevent gender-stereotyping must be subtly gender-normative, and reinforce certain norms so as to preempt any outputs that frame the atypical as abnormal. Most systems tacitly incorporate a model of their audience, after a fashion, in their objective functions, in their aesthetic criteria, and in the ways they frame their outputs, but those audience models must increasingly integrate a clear sense of what affronts as much as what delights. In fact, a robust super-ego is essential for systems that adapt to their users to grow their spaces or to evolve new objective functions, so that those systems are not corrupted over time, or worse, weaponized by their most malicious users to attack others.

In this position paper we hope to stimulate debate within the computational creativity community, by examining the responsibilities that we all bear as system builders, that our systems bear as meaning makers, and that the larger public bears as potential end-users and informed critics. The issues are much larger than any single paper or project can hope to address, but we can start by confronting the assumptions that underpin our conceptions of offense, before outlining some strategies for mitigating its most insidious forms.

Epic Fails

Even in the age of "going viral," a single nugget of content cannot make a career, but a single tweet can still ruin one. Consider the case of comedienne Roseanne Barr who, back in 2018, led the revival of her eponymous sitcom. The new show was feted by critics, but came to a crashing halt after Barr tweeted the following in the early hours of May 29:

muslim brotherhood & planet of the apes had a baby=vj

The "vj" of her tweet refers to Valerie Jarrett, an appointee of the Obama administration *and* a women of colour. While Barr's conceit is a technically deft conceptual blend of two very different input spaces (Fauconnier & Turner, 2002), it draws on an odious animal trope long favoured by racists. It is not enough to judge blends on technical grounds alone; unlike Barr, we (and our systems) cannot be so charmed by a novel conceit that we are blinded to its intrinsic ugliness. Barr soon regretted her tweet, but after-the-fact evaluation often comes too late, and at a high price. Barr was quickly fired by her network from a show that once bore her name.

Barr undoubtedly recognized her own use of this trope, but did not consider it ugly until her career was threatened. We need our generative systems to do both of these things at the same time: to recognize the tropes that seem to apply to their outputs, at least in the minds of an audience, and to recognize the potential for harm contained within them. A failure on each of these fronts had been the undoing in 2016 of a flagship Twitterbot by Microsoft, called @*TayAndYou*. "Tay," designed as a conversational chatbot, was given the language model of a curious teenager. Although this model was carefully stocked with anodyne content, Tay was also designed to learn from those it interacted with, and to adopt stances in its tweets rather than simply sit on the fence. Tay would prove that inner regulation is a good starting point for an interactive bot, but no amount of curated content for hot-button issues – Tay had upbeat views on Caitlin Jenner ("a hero & a stunning beautiful woman") and sympathetic views on recent victims of police violence – could stop its generative model being overwhelmed by malign influences. It was soon parroting very different takes on these topics:

caitlin jenner pretty much put lgbt back a 100 years
as he is doing to real women

Despite Microsoft's best intentions, Tay was a signal failure of outer regulation in a bot. Even a naïve filter would have found Tay's repeated use of ethnic slurs and racial epithets offensive, and identified topics of great sensitivity where a bot like this should never dare to tread. Tay dared, however, and was soon denying one genocide (the Holocaust) while advocating another (of minorities in the United States). Barr compared a black Obama appointee to an ape in her tweet, but Tay would describe Obama himself as a monkey, and – in the same tweet – accuse George W. Bush of planning the 9/11 attacks. Microsoft was forced to euthanize its bot less than 24 hours of it going live on Twitter, much as the ABC television network was later moved to cancel Barr.

Microsoft blamed the bot's rapid slide to the dark side on "a coordinated attack by a subset of people [that] exploited a vulnerability in Tay" (Ohlheiser, 2016). This vulnerability was not a secret backdoor or a code glitch, but a gaping hole in its design. Microsoft failed to provide its bot with even the most rudimentary outer regulator, relying instead on the kindness of strangers to treat the bot with gentle goodwill.

Little Troublemakers

Offense can be situated on the orthogonal axes of content and behaviour. On Twitter, simple bots show that it needn't take much of either to make a mark. Yet as our generative systems push their resources to the limit, they exemplify the old saw that "a little knowledge is a dangerous thing."

A bot intending to provoke can lean on its content or its behavior. Generally, the more provocative the behavior, the more benign the content can be and still cause offense. The converse is also true, since malign content does not require malign behavior to make it offensive. Consider the case of *@StealthMountain*, a Twitterbot that knows just one thing, how to spell "sneak peek," and can do just one thing, search for Twitter users who misspell this phrase as "sneak peak" (note the sympathetic choice of "peak" instead of "peek") so as to target them with the solicitous message "I think you mean 'sneak peek'." Although the mistake is minor and the advice benign, few of the bot's targets take kindly to these intrusions. Rather, this little bot can provoke some extreme reactions from those who decry its actions as the work of a "busybody," a "spelling fascist," or "the grammar police." This is the bot's larger purpose: to entertain others with the oversized reactions of those offended by its tiny intrusions.

Benign content is not always welcome content, but this is what it means to be an intrusive generator. The offense that is inflicted by such intrusions is compounded when content is deliberately pushed at those who are least likely to enjoy it. Take, for instance, the behaviour of *@EnjoyTheMovie*, a bot that is more targeted in its intrusions and more varied in its use of factual knowledge than *@StealthMountain*. The bot's knowledge-base comprises a mapping of movie titles to movie spoilers – key elements of a plot that are ruined if revealed in advance – which it targets at those who express an interest in watching those movies for the first time. The bot shows that timing is an important part of offense, since facts only become spoilers if uttered at the wrong time. The bot can afford to be scattershot in its targeting, for although it cannot accurately assess whether a tweet implies that its author is oblivious or not to a given spoiler, potential targets are plentiful on Twitter, and some are sure to be offended.

It is not a coincidence that each of these bots has been suspended by Twitter, since its policies frown just as much on unwelcome behaviors as unwelcome content. A "model" bot does not target unsolicited content at others, but creates content that will lead others to seek it out for themselves. A case in point is Darius Kazemi's *@twoheadlines* bot, which generates weird and sometimes wonderful cut-ups of news headlines. The bot tweets its novel mashups – generated by swapping a named entity in one headline for one found in another – into its own timeline, and those of its followers. Some cut-ups are plausible, while some rise to the level of a humorous blend, but many more are just plain odd. In this cherry-picked pair, one is plausible, the other darkly comic:

President Trump Trying to Bring Nintendo Entertainment System Back to Life

Miss Universe attacks North east Nigerian city; dozens killed

When *@twoheadlines* lifts a named entity from its home domain and transplants it to the alien setting of a headline in which it seems just about possible, but still highly unusual, then humour is a likely by-product. In these new settings, famous rappers can win the Superbowl, or a movie star can have a closing down sale, or a US senator can "open up to third party developers." While the bot is as scattershot as its simple cut-up approach would suggest, its sporadic flashes of accidental wit gain it followers while keeping it on the right side of Twitter's code of bot conduct. In fact, because the bot splices none of its own content into its outputs, and relies solely on the words and entities that it finds in actual headlines, it has a built-in inner-regulator by default. Since it applies the cut-up technique to real headlines, which are themselves the products of inner- and outer-regulation by editors and reporters, *@twoheadlines* never uses terms that would seem out of place in a family newspaper.

But inner regulation offers no protection against the kind of offense that emerges from the combination, not the parts. The Franken-headlines of *@twoheadlines* certainly fall into this possibility space. Instead of a politician "opening up to third-party developers" – a cut-up that seems to satirize the role of money in politics – imagine that the bot had instead spliced a celebrity such as Caitlin Jenner or Angelina Jolie. The bot sets out to spark incongruity with its cut-ups, so that some incongruities will rise to the level of satire and farce. Yet it lacks any ability to appreciate its own products, or to predict who will be the ultimate butt of the joke. So, will its humorous effect be restricted to a specific named entity, or might some see it as a broadside against a class of entities? Kazemi has struggled with these possibilities (Jeong, 2016), and his *@twoheadlines* is a model bot in other respects too. He is especially concerned by the possibility of unintended slights, in which meaning transcends the specific to target a large group of people, from communities to ethnicities. His ounce of prevention takes the form of an outer regulator.

In particular, Kazemi is concerned by the propensity of the cut-up method to cross gender boundaries and generate headlines that seem to sneer at the transgender community. He offers as an example two input headlines: one contains "Bruce Willis," and the other a female actor who "looked stunning in her red carpet dress." Although *@twoheadlines* might elicit laughs and likes with the cut-up "Bruce Willis looked stunning in her red carpet dress" – since the image it paints is so vivid an example of a comedy trope–it might also reinforce the validity of those old-school tropes. And while the bot's slight is without malice, those who retweet it far beyond its intended audience may not be so innocent. To deny them this opportunity, Kazemi imposes a gender regulator on his bot's traffic in named entities. When one named individual is swapped out for another, this regulator requires that each has the same gender. So, while Joe Biden can replace Donald Trump, and Meryl Streep can replace Julia Child, Robin Willians cannot replace Mrs. Doubtfire.

This is the only aspect of *@twoheadlines*'s behaviour that is informed by its own small view of the world. An outer regulator needs knowledge, to tell it how and when to act. Yet the bot's use of knowledge will strike many as ironic, since it enforces gender-normativity at the generative level to disallow unwanted heterogeneity at the surface. It is not that such heterogeneity is undesirable, rather that it cannot be safely controlled, nor can its consequences be predicted. Although this applies to all of his bot's outputs, Kazemi has chosen to restrict its freedoms in one particular area where he feels this lack of control can lead to the worst offenses. In effect, he has adopted a *homeopathic* approach to outer regulation, by integrating into his system a tiny element of the peril he hopes to protect it – and its audience – against.

The Bot Police

As reflected in the common fate of *@StealthMountain* and *@EnjoyTheMovie*, bots that become a nuisance are quickly suspended once irate users report their offenses to Twitter. This user-led policing of bots is mirrored by Twitter's own automated efforts to weed out bot abuses on their platform. Although these efforts are rather unsophisticated, and focus more on overt offense than subtle manipulation, they should give pause to the creators of all generators on social media.

The principle of *caveat emptor* governs the use of vulgar and offensive language on social media, for on platforms designed to connect people, *what* you say is often no more important – and sometimes less so – than *who* you say it to. It is the coupling of content and behavior that Twitter aims to police, which is why frivolous non-vulgar bots like those above have so short a life-span on the platform. Consider how Twitter reacts to the following tweet from a bot that posts personalized colour metaphors for famous users. The bot also creates images, but here we consider the text only:

> **I painted "wise-cracking Jar-Jar Binks" from @~~anonymized~~'s tweets, with goofy redneck-red, foolish ass-brown and laid-back Lebowski-weed-green.**

This tweet, from a bot named *@BotOnBotAction*, offers a number of reasons to feel mildly offended. The word "ass," meaning "donkey," is also a mild anatomical insult; "weed" can also mean an illicit drug, as it does here; and "redneck" (meaning "oaf") is now a politically-charged term. None of these words is offensive in itself, and there is no shortage of uncontroversial tweets with some or all of them on Twitter. The tweet still earned Twitter's ire, prompting this response:

> **Violating our rules against hateful conduct:**
>
> **You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.**

Twitter shrewdly omits a word-specific rationale as to why this tweet earns the bot a one-week suspension. Dictionary-based models of offense detection, as we discuss next, are easily circumvented if abusers know which words to avoid. This explains the spammer's love of "pen1s" and "v1agra," but no single word seems to trip Twitter's silent alarm here. Rather, it appears to be a combination of mildly suggestive terms that might be intended as insults with the @-mention of another user that triggers the platform's intervention.

Dictionary-Based Approaches

Many comedians take delight in veering as close to the line of offensiveness as possible. Others actively cross this line, if only to show how arbitrarily drawn it sometimes seems. A stand-up routine by the comedian George Carlin in 1972, recorded for posterity on his album 'Class Clown', riffed on the taboos of the day and gave us "the seven words you can never say on TV." Though Carlin's list had no factual basis in regulatory standards – it was likely chosen to mirror the Bible's seven deadly sins – it struck a chord with viewers:

> "shit, piss, fuck, cunt, cocksucker, motherfucker and tits"

Carlin critiqued the rigidity of the list, noting that its entries were not all equally offensive. He also noted the absence of many other, somewhat milder terms, speculating that their omission was related to their "two-way" ambiguity. A great many of the words that we deem offensive have legitimate uses too, making their inclusion on a blacklist problematic.

The prolific bot-builder Darius Kazemi provides a range of resources for budding developers, from quirky data-sets to bolt-on modules such as *WordFilter* (Kazemi, 2015). The latter is a blacklist-based outer regulator that proscribes ten times as many words as Carlin's original list. In addition to multiple variations of the N-word, his list includes a range of coarse sexual terms, and those that denigrate others on the basis of race, gender, body type and mental deficiency. His list is not without its quirks, however, and proscribes terms such as "idiot," "crazy," "dumb" and "lunatic." These may well be hurtful, but they are not inherently offensive.

WordFilter pounces on any word contained on its list, or on any word that contains an entry as a substring. Because it over-generates by design, it matches true positives – such as *dickpix* and *bitchslap* – that use its entries as morphemes, and many false positives too, such as *snigger*, *homology* or *Scunthorpe*, that are utterly unrelated. Over-proscription by a blacklist causes under-generation in the regulated system, but when the space explored by a generator is already vast, an outer regulator can easily afford to be less than surgical. Nonetheless, an overly-general blacklist suggests a role for a complementary "white" list that enumerates any common exceptions to substring matching, such as "sauer*kraut*."

Notably, Kazemi's list does not include the scatological nouns and sexual verbs that make up what we traditionally think of as "blue" or vulgar language, since base vulgarity is not in itself offensive. Its aim is not to regulate bad taste but to minimize the possibility of accidental hate speech, although *Wordfilter* will still fail to flag outputs of the form "all [ethnicity] are [vulgarity]." While a system must strive to avoid clear signs of hateful intent, offense is contextual, and arises from the whole rather than from any single part.

WordFilter's contents are a mix of the not always good, the frequently bad and the unambiguously ugly. Words that should never be used in polite discourse sit cheek-by-jowl with words that only become offensive in specific contexts. To *WordFilter*, however, they are all equally abhorrent. A more nuanced lexical approach to offense can be found in resources such as *HateBase.org* (Keating, 2013), an online resource that is indexed by geography and severity, and in reports commissioned by national broadcasters to determine the community standards by which they should abide. The 2016 report of the UK regulator *Ofcom* (Ipsos Mori, 2016) is typical of the latter. It distinguishes general swear words from discriminatory language, identifies lexical innovations in the latter, and surveys the acceptability of different terms to the viewing/listening public at different broadcast times. Each is a rich source of data in which system builders can find the lexical material for their blacklists, and – if shades of offense are to be gradated – their *grey* and *white* lists too. In principle, HateBase's atlas of "what offends where" can allow a regulator to tailor its filter to the norms of a region, to accept words in one setting that it might avoid in another. However, if harvesting external sources such as these, one must accept the subjectivity of their authors, as when, e.g., HateBase deems "kraut" to be just as offensive as "Nazi."

Dictionary-based regulators are susceptible to dictionary-based attacks. Consider a social-media campaign by *Coca-Cola* that ran in 2016. Consumers were invited to attach an upbeat, on-brand text to an animated GIF so that the pairing might then "go viral." The company employed a word filter to regulate the kinds of text that mischievous users might try to link with the *Coca-Cola* brand, so that the campaign would not become a vector for politics or hate speech. To estimate the size of the company's blacklist, Bogost (2016) ran an English dictionary through the app, noting the words that caused it to balk. He was surprised both by the number and the kinds of words on its blacklist, from 'capitalism' to 'igloo' to 'taco.' While few of its entries were offensive in isolation, many more might serve as the building blocks of a cultural critique or a racist attack. When the reputation of a company or a product is protected with a blacklist, a great many innocent words must necessarily become suspect.

In an earlier misstep in 2015, Coca Cola had encouraged consumers to append the hashtag *#MakeItHappy* to tweets with a less than happy tenor, so that a Twitterbot might then rework each one as a cute piece of ASCII art. The campaign was soon undone by yet another bot that attached the tag to a stream of dull extracts from Hitler's *Mein Kampf*, thereby duping the company into stamping its brand onto an odious work (Read, 2015). Ultimately, a blacklist is an uncreative solution to a creative problem, and generative systems give hostages to fortune whenever they elicit inputs from others who are themselves impishly – or even wickedly – creative.

No blacklist, however broad, would catch the following ill-advised use of a prejudicial stereotype by a Twitterbot:

On the anger theme, @~~anonymized~~, I only became as emotional as a woman after I read "Hamlet" by William Shakespeare.

This tweet was generated by a book recommendation bot with a figurative turn of phrase, called *@ReadMeLikeABot*. As described in Veale (2019), the bot harvests its stock of similes–much like its books–from the web, and while it can distinguish similes from comparisons, and tell sincere from ironic cases, it cannot identify those that are carriers of bias. In the following tweet, its blacklist is blind to another gaffe:

On the mothers theme, @~~anonymized~~, I used to be as charming as a photo album of the Holocaust until I read "The Bone Setter's Daughter" by @AmyTan.

Words such as "woman" and "Holocaust" do not belong on a blacklist, and should not be scrubbed from a search space by inner regulation, but do need to be used with some care. The problem here is compounded by irony, which masks an implied negative with an overtly positive frame. A solution, of sorts, is to employ a "red" list of sensitive terms that can not be used in a hostile or ironic setting, and that perhaps meet a higher confidence threshold whenever they are used.

Corruption and Weaponization

As our generative systems migrate from the lab to the web, the potential to be corrupted and weaponized by malicious third parties grows considerably, but the consequences of a misstep can be so much worse if it is made on social media. Broadly speaking, there are two ways in which offense can be weaponized on such platforms. In the first, a machine's offensive outputs are spread by those aiming to widen their impact. By causing them "go viral," offenses can be spread far beyond the system's immediate users. In the second, a learning system's judgments might be subverted over time by exposure to bad examples that, if repeated often enough, cause it to adapt its knowledge-base to this new normal. A system that is corrupted in this way may be led to generate offensive statements about a third party, or even to address the offense to that party directly on a social media platform.

Let's look again at Darius Kazemi's bot *@twoheadlines*, and recall his use of an outer-regulator to disallow any cut-ups that cross gender lines. This bot is effortlessly prolific, since new material for its cut-ups is constantly produced by the news media. So it can afford to designate a subset of its possibility space as forbidden territory. But this regulator is also a generator in reverse, as are so many outer regulators, because it can be inverted to generate that which it seeks to prevent. For if *@twoheadlines* actively forced its cut-ups to cross the line, and only swap entities of different gender, it would open many more opportunities for offense. While the likelihood of a responsible designer pursuing this option is low, the problem has a more insidious variant. Suppose that a generator has the capacity to learn from its user-base, and to adapt its generative mechanisms to their preferences. If users up-vote examples of the kind of output that an outer-regulator should be designed to throttle, this generator may eventually learn to produce *only* that kind of output. The up-voting process can itself by automated, by a malicious third party aiming to subvert the choices of a generator. We can, in this sense, view an outer-regulator as the immune system of an adaptive generator. Much as the immune system of a biological agent must discriminate between *self* and *other*, to prevent the self from being corrupted, a regulator must preserve the values that make up a generator's digital self.

It seems clear from the Tay debacle that Microsoft gave its eager learner an immune system with very few defenses. Bender *et al*. (2021) caution that web-trained models can act as "stochastic parrots," and tellingly, the most corrosive assaults on the bot's language model were prefixed "repeat after me." While such parrots are facile in the production of surface forms, they fail to understand what they generate, just as they fail to grasp what it is they are learning. Because Twitter handles are just another form of content, it was not long before Tay learned to tweet collateral abuse at specific targets, as when it assailed a vocal games creator and critic with: "@UnburntWitch aka Zoe Quinn is a Stupid Whore." Personal identifiers may look like any other kind of text to a language model, but they should never be treated as such. Blacklists at least recognize that not all symbols are equal, but our systems need special policies for special signifiers.

Accidental Chauvinists

It is now widely accepted that generative models which are trained on web data are prey to all of the biases, prejudices and illiberal stereotypes that the web has to offer. Moreover, a larger training set is not necessarily a more diverse one, especially if it is pre-filtered to remove non-normative data. As observed in Schlesinger *et al*. (2018) and Bender *et al*. (2021), these filtering and data-cleaning efforts can further marginalize under-represented communities, and reinforce dominant, if unmarked, norms of maleness and whiteness. But a generator need not be prejudiced by its training data to show an apparent bias, and we must distinguish between bias-free and bias-blind generation. Even systems that make purely random decisions in a uniform possibility space are susceptible to the appearance of in-built bias if they lack an awareness of how their choices might be viewed by others.

Consider the story-generation system described in Veale (2017). This generator is wholly symbolic in operation, and has no training data from which to absorb a biased outlook. It inserts well-known characters from fact and fiction into plots that are assembled by its story grammar, and renders the resulting blend as a narrative that draws vivid elements of *mise en scène* from its detailed character representations. Although those details include gender and political stance, no gender-normativity is enforced when filling a story role. Rather, the generator seeks to produce its sparks elsewhere, in the pairing of characters that seem well-suited and oddly inappropriate at the same time. So it may, for instance, pair Steve Jobs and Tony Stark, Ada Lovelace and Alan Turing, or Hillary Clinton and Donald Trump in a mad love affair. It knows enough to be provocative, but not enough to grasp the full implications of its provocations. When it pairs Luke Skywalker to Princess Leia in its romantic retelling of *Star Wars*, it does not know that its lovers are brother and sister. Its generative reach exceeds its generative grasp by design, in ways that invite Eliza effects but avoid Tale-spin effects.

However, not every Eliza effect is a desirable effect, and the system's gender-blindness sometimes leads it to create narratives that appear as the products of systemic prejudice. Audiences suspend disbelief when Luke woos Leia, but are less forgiving when Leia reciprocates by cooking for Luke. The former can be chalked up to a lack of film knowledge, but the latter is more readily attributed to sexist stereotypes. In our experience, audiences make one-shot judgments as to in-built biases when those biases are prevalent in society.

A variant of the same system (Wicke and Veale, 2021) elicited another one-shot determination of bias. This paper provides a video recording of robots enacting an imaginary romance between Hillary Clinton and Donald Trump. The plot, as generated by the story grammar, calls for Donald to propose to Hillary, and for Hillary to accept. The rendering of the tale as an English narrative then seeks out a post-hoc rationale for her acceptance, by searching its character data for the positive qualities that might make Trump desirable. It chooses from among these qualities at random, and picks *rich* and *famous* as those which attract Hillary to her suitor. These, however, conform to the gold-digger trope, and the story was deemed innately sexist by the paper's reviewers. The paper was accepted only once the story and video were altered, and a "shepherd" had confirmed their lack of bias.

People's sensitivities in this area are well-founded, even if their assessments lack rigour and a clear pattern of bias. Systemic bias in society makes one-shot judgments of our creative systems more likely, and more reasonable, if those systems deal with culturally-freighted signs or concepts. It is not enough that the system above is bias-free by design, because it is also blind to bias by default. So, what would it mean for a system to be free of bias *and* bias-aware? Storytelling systems already consider audience reactions to the decisions they make as they relate to character and plot, so the analysis of perceived bias is just another consideration.

It is important that our systems do not overreact, by inner regulating their search spaces to prune perfectly acceptable possibilities – such as a woman cooking a meal, or indeed, anyone ever cooking anything – from being explored. We have more tools at our disposal than filtering. For example, Wicke and Veale (2021) describe stories at multiple levels of enactment: the actions of performers that act out in-story character roles, the actions of omniscient narrators, and the actions of actors as self-conscious performers. As such, a bias-aware storyteller can preempt the perception of bias by weaving meta-level commentary into its performance – e.g., "Why can't *Luke* cook?" – that signal its awareness of illiberal stereotypes and its willingness to call them out. An outer regulator need not be a censor. it can also be a voice of moderation that explains, critiques, and educates too.

Caveat Pre-Emptor: A Practical Manifesto

Automated solutions to the mitigation of generative offense will, if they are practical, reflect the maxim that the perfect is the enemy of the good. For no system with the expressive power to pack interesting ideas into diverse forms will ever be able to prevent all kinds of offense, real or imagined. So we offer this practical manifesto for mitigating offense with the caveat that it is far from perfect. However, an inability to provide sufficient solutions should not prevent us from exploring necessary solutions, partial though they may be.

Blacklists are necessary, but they are far from sufficient

A blacklist offers a crude solution to offense, but it is still a vital component of any generative system that manipulates cultural signifiers. A blacklist is the ultimate firewall, a last line of defense that recognizes the need for further defenses. Whether those added layers are symbolic, statistical, neural or hybrid, a blacklist must still sit at the bottom of the stack.

Some words are offenses, but others only facilitate offense

We should not overload a blacklist with signs that facilitate ugly behaviours but which are not in themselves offensive. Words that demand great sensitivity, such as "Holocaust," or legitimate signifiers of ethnic and racial identity, should not be listed as offenses just because a generator lacks the means to adequately predict what they signify in context. If necessary, such terms can be placed on other 'watchlists' that do not stigmatize their presence (e.g. a *grey* or *red* list).

There are policies for signs that have predictable behaviours

If different kinds of signifier entail different behaviours in an audience or a transport layer (e.g., Twitter), a generator should define a policy for handling each kind of behaviour. A policy can be as basic as the filtering of @-mentions in a system's outputs, to avoid redirecting offense at collateral parties, or the total avoidance of hashtags, so that a system is not co-opted onto another's bandwagon or social cause.

Outer Regulation is always preferable to Inner Regulation

Concepts and signs that are denied to a generative system due to inner regulation should still be mirrored in the outer regulatory layer, so that the system *knows* what it is denied, and does not acquire them through other channels. Outer regulation supports explanation; inner regulation does not.

Every "creative" system has an outer-regulator by default

Systems that generate what their rules will allow, without a subsequent layer of self-evaluation, are "merely generative" (Ventura, 2016). Since creativity tempers generativity with self-criticism and a willingness to filter whatever drops off the assembly line, a creative system is defined by its outer regulator. This regulator's remit is typically aesthetic, but it might also include any ethical concerns that can be codified (Ventura and Gates, 2017). So a system that unknowingly generates offense, and that lacks mechanisms to critique its own outputs, can hardly be deemed "creative," no matter its achievements along other, more aesthetical dimensions.

Blacklists should not be replaced with black box regulators

When blacklists are public, offenders know to expend their creative energies elsewhere. When secret, they are closed to scrutiny but still vulnerable to dictionary-based attacks. A self-regulating generator should be capable of explaining its decisions (Guckelsberger *et al*, 2017), since accountability built on transparency can make its mistakes more tolerable.

A creative system needs a moral imagination too

Blacklists and similar approaches "tell" a generator what is out of bounds, but do not explain their curbs, or provide the means for the generator to make the same choices for itself. A blacklist is no substitute for a moral imagination, but the latter requires experience of the world. However, new data sets (Forbes *et al*., 2020; Lourie *et al*., 2020) annotate real-word scenarios with a normative ethics, sifted and weighted by competing perspectives. A moral imagination trained on these data sets can, in principle, derive and explain its own acceptability judgments. Moral justification is just another kind of framing, but one that must be adequately resourced.

When it comes to offense, learning should narrow the mind

In systems that learn from feedback, a system can add to its blacklist (or equivalent filter) but never take from it. A user may trick a system into blacklisting an acceptable term, but not trick it into relabeling a blacklisted term as acceptable.

Social attitudes evolve quickly, and our regulators must too

Shifts in attitudes can be sudden and disruptive, not steady and continuous. What was widely acceptable just a decade ago may now be subject to severe reappraisal, as in e.g. the books of Dr. Seuss and other authors whose stereotypical depiction of minorities is now seen as insensitive. We need agile regulators that can shift just as quickly, and that can retrospectively filter past mistakes in their back catalogues. For instance, a Twitterbot may periodically delete old posts that no longer pass muster with its evolved outer regulator.

Homeopathy works, but only in small doses

Creativity is always a trade-off: between novelty and value, or generative reach and generative grasp, or principle and practice. A system seeking to avoid offense may inoculate itself against the very worst by embracing a smaller dose of that which ails it, but accountability demands that we make our trade-offs public so that their costs can be appreciated.

The sharing of imperfect solutions supports standardization

A cynic might see the public sharing of our system's filters as an act of virtue signalling, but sharing enables standardization, and standardization promotes compliance. To foster public trust in creative systems, we can agree on standards-backed compliance levels that users can rely upon to make their own decisions, or to understand those of our systems.

Conclusions: We All Bake Mistakes

Our community has traditionally focused on the building of *ingenious ingenues*, that is, on creators that are as naïve as they are clever. While the link from ingenuity to creativity is an obvious one, naiveté is also a predictable consequence of creators that generate more than they can appreciate. A semiotic system has the capacity to naively offend when its signs can reference the larger culture in which it is situated. Because language offers this referential capacity in spades, our principal focus here has been on linguistic generation, but other modes of creative generation are also susceptible.

Tay, for instance, offended with images as well as words, as when it tweeted images of Adolf Hitler labelled "Swag." For multimodal generators, clear offenses in one modality may cause innocent choices in another to become suspect. Since a potential for offense attaches to all but the simplest generators, it is a problem that we must all tackle together. This will require more than codes of practice, such as those adopted by many bot builders (Veale and Cook, 2016), but goal- and task-specific regulators of offense that allow our systems to reason about their own likely transgressions.

The mitigation of offense calls for an awareness of others and of the self. A generator must be aware of its own goals, of what it wants to express and how it plans to express it. It must also be wary of how others interpret what it does, and avoid misconceptions both accidental and deliberate, since missteps that are embarrassing in isolation can be magnified to viral dimensions with the help of malicious actors. Inner regulation is clearly not conducive to this self-awareness, since it blinds a system to the iniquities it wishes to prevent. Inner-regulation makes offense an "*unknown unknown*," so that a system lacks the knowledge to cause it deliberately, and crucially lacks the tools to diagnose it in others. Outer regulation, in contrast, models offense as a "*known known,*" even if regulators know only enough to shun certain signs. In this view, offense mitigation should not be handled as a bolt-on module – a generative septic tank – but as one part of a broader effort to make our systems more self-aware.

In practice, many systems will use both inner and outer regulation, tacitly or otherwise. Systems that are trained on selective datasets, curated to exclude all undesirable cases, are tacitly designed to be ignorant of the causes of offense. Systems like these still require an outer regulator to police emergent offense, or to filter new learning materials as they are acquired through interactions with teachers and users. When a creative system is capable of producing hateful or offensive content, however accidentally, regulators must be just as creative in their ability to appreciate its impact. For if we, as system builders, are willing to take the credit for a system's fortuitous accidents, when it seems to transcend its programming and show wisdom beyond its design, we must also be ready to take the blame for its humiliating missteps.

References

Veale, T. (2017). Déjà Vu All Over Again: On the Creative Value of Familiar Elements in the Telling of Original Tales. In *Proc. of ICCC 2017, the 8th International Conference on Computational Creativity*, Atlanta, Georgia, June 19-23.

Bender, E.M., Timnit, G., McMillan-Major, A. and Shmitchell. S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proc. of FAccT '21, the Conference on Fairness, Accountability, and Transparency*, March 3–10.

Bogost, I. (2016). Things You Can't Talk About in a Coca-cola Ad. *The Atlantic Magazine*, January 28 edition.

Charnley, J.W., Pease, A. and Colton, S. (2012). On the Notion of Framing in Computational Creativity. In *Proc. of the 3rd International Conference on Computational Creativity*, 77-81, Dublin.

Fauconnier, G. and Turner, M. (2002). *The Way We Think. Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Forbes, M., Hwang, J., Schwartz, V., Sap, M. and Choi, Y. (2020). Social Chemistry 101: Learning to Reason about Social and Moral Norms. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 653–670.

Guckelsberger, C., Salge, and C., Colton, S. (2017). Addressing the 'Why?' in Computational Creativity: A Non-Anthropocentric, Minimal Model of Intentional Creative Agency. In *Proc. of the 8th International Conference on Computational Creativity*.

Ipsos Mori (2016). Attitudes to potentially offensive language and gestures on TV and radio. UK: *Ofcom*, https://bit.ly/3qvIBQd

Jeong, S. (2016). How To Make A Bot That Isn't Racist: What Microsoft could have learned from veteran botmakers on Twitter. *Motherboard (Tech By Vice),* March 25 online edition.

Kazemi, D. (2013). WordFilter. github.com/dariusk/wordfilter

Keating, J. (2013). Mapping hate speech to predict ethnic violence. *Foreign Policy*, April issue.

Lourie, N., Le Bras, R., and Choi, Y. (2020). Scruples: A Corpus of Community Ethical Judgments on 32000 Real-Life Anecdotes. *arXiv*:2008.09094.

Meehan, J.R. (1977). TALE-SPIN, an interactive program that writes stories. In *Proc. of the 5th International Joint Conference on Artificial intelligence*, 91–98. Cambridge, Massachusetts..

Ohlheiser, A. (2016). Trolls turn Tay, Microsoft's fun millennial AI bot, into a genocidal maniac. Washington Post, March 25.

Read, M. (2015). Make Hitler Happy: The Beginning of *Mein Kampf*, as Told by Coca-Cola. *Gawker*, https://bit.ly/3bB18X7

Ritchie, G. (2007). Some Empirical Criteria for Attributing Creativity to a Computer Program. *Minds & Machines* **17,** 67–99.

Schlesinger, A., O'Hara, K.P. and Taylor, A.S. (2018). Let's Talk About Race: Identity, Chatbots, and AI. In *Proc. of the CHI Conf. on Human Factors in Computing Systems*, 1–14. Glasgow, UK.

Veale, T. and Cook, M. (2016). Twitterbots: Making Machines that Make Meaning. Cambridge, MA: MIT Press.

Veale, T. (2019). Read Me Like A Book: Lessons in Affective, Topical and Personalized Computational Creativity. In *Proc. of the 10th International Conference on Computational Creativity,*

Ventura, D. (2016). Mere Generation: Essential Barometer or Dated Concept? In *Proc. of the 7th International Conference on Computational Creativity*.

Ventura, D. and Gates, D. (2018). Ethics as Aesthetic: A Computational Creativity Approach to Ethical Behavior. In *Proc. of the 9th International Conference on Computational Creativity*.

Wardrip-Fruin, N. (2009). *Expressive Processing: Digital Fictions, Computer Games, and Software Studies*. Cambridge, Massachusetts: MIT Press.

Weizenbaum, J. (1966). ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9, no. 1 (Jan.): 36.

Wicke, P. and Veale, T. (2021). Are You Not Entertained? Computational Storytelling with Non-Verbal Interaction. In *Proc. of HRI'21 Companion*, *the ACM/IEEE International Conference on Human-Robot Interaction,* Boulder, Colorado, March 8-11.