How Twitter Gamifies Communication

C. Thi Nguyen

Forthcoming in *Applied Epistemology*, ed. Jennifer Lackey (OUP)

Twitter is now one of our primary venues for public discourse. But it is not a neutral or transparent medium. Twitter shapes how we interact, who we interact with, and — perhaps most importantly — it suggests specific goals for those interactions. Twitter doesn't just provide a speaking platform, nor are its effects confined to algorithmic filtering. Twitter shapes our goals for discourse by making conversation something like a game. Twitter scores our conversation. And it does so, not in terms of our own particular and rich purposes for communication, but in terms of its own pre-loaded, painfully thin metrics: Likes, Retweets, and Follower counts. And if we take up Twitter's invitation and internalize those evaluations, we will be thinning out and simplifying our own goals for communication.

Let's take a step back. Twitter is at once pluralistic in its scope and monolithic in its technological form. Twitter is pluralistic because it offers relatively open access to powerful resources for public discourse. Anybody can form an account, and anybody with the right feel for the medium, it seems, can gather enormous numbers of followers. Twitter democratizes access to large-scale communication, which once had been held by a relatively small number of media companies.[1] At the same time, Twitter is monolithic, because everybody who uses

[1] My understanding of Twitter here has been particularly informed by Zeynep Tufekci's (2017) thoughtful analysis of the advantages and disadvantages of this pluralism.

Twitter must communicate through the same interfaces, and be subject to the same algorithms.[2] What is the impact of so much public discourse being shoveled through one platform?

Other discussions of Twitter have focused on the enforced shortness of tweets, the influence of hidden algorithmic filtering, the promotion of group polarization, the lack of accountability mechanisms, and the collapse of conversational contexts (Sunstein, 2009, 46-96; Marwick and boyd, 2011; Miller and Record, 2013; Frost-Arnold, 2014; Rini, 2017). I would like to focus on another basic feature of Twitter — one whose importance and impacts, I think, has not been adequately appreciated. Twitter gamifies communication by offering immediate, vivid, and quantified evaluations of one's conversational success. Twitter offers us *points* for discourse; it *scores* our communication. And these game-like features are responsible for much of Twitter's psychological wallop. Twitter is addictive, in part, because it feels so good to watch those numbers go up and up. In fact, the design of Twitter and its scoring mechanisms have been significantly informed by design strategies fostered in the Las Vegas gambling industry — strategies which overtly seek to increase the addictiveness of their products.[3]

The clear scoring system brings with it another very game-like aspect: a clear and unambiguous ranking. We usually don't emerge from the party with a ranked list of who the best

[2] This pattern of thought has been particularly influenced by Tufekci's (2018) discussion of how the Internet era has democratized communication, but at the same time subjected all online communication to a few very small portals. The Internet is democratic because anybody can put up a web-page, but it is monolithic because we all find web-pages using Google Search - so Google Search's algorithm becomes an all-powerful control on our collective attention. Obviously, in the background, is Marshall McLuhan's (McLuhan, 1964) discussion of the impact of medium over content.

[3] Natasha Dow Schull's *Addiction by Design* (2012) offers a thorough look at the technological innovations of the gambling industry to optimize the addictiveness of their products. Since that book, Schull has been vocal about how those technologies have been adopted by gaming and social media companies (Madrigal, 2013; National Public Radio, 2014; Seymour, 2019).

conversationalists were. Twitter, on the other hand, offers both short-term rankings (Likes and Retweet numbers for each tweet) and long-term rankings (Follower counts). Most importantly, the rankings are entirely unambiguous. Unlike conversation in the wild, I can know exactly how well each tweet did, and I can instantly compare my overall popularity with that of any other user. This can provide all sorts of pleasures: the thrill of victory, when we see those numbers tick up; and the sense of long-term achievement, presented in precise and unquestionable quantitative form.

Supporters of gamification say that it is a technology for increasing motivation. Gamification can supposedly imbue everyday activities with all the fun and excitement of a game.[4] Here, then, is an optimistic view of Twitter: by gamifying public discourse, Twitter increases overall participation, and so helps us to reap the rewards of public discourse — such as a more fully politically engaged populace.

I do not accept the optimistic view. Crucially, I don't think that gamification merely increases our motivation to perform an activity while preserving all the original goods of that activity. Gamification increases our motivation by changing the nature of the activity. Often, the goals of ordinary activity are rich and subtle. When we gamify these activities, we change those goals to make them artificially clear. Games are more satisfying than ordinary life precisely because game-goals are simpler, clearer, and easier to apply. In games proper, this simplification isn't particularly problematic, because the goals are peculiarly artificial. Game activities, and their associated goals, are usually kept secluded from ordinary life. But there

[4] This point was put most influentially by Jane McGonigal (2011). For critical discussion, an excellent starting place is (Walz et. al., 2015).

is no such protective separation when we gamify ordinary activities. To reap the motivational benefits of gamification, we must re-shape the ends which govern our real-life activities.

Pre-gamification, the aims of discourse are complex and many. Some of us want to transmit information or to persuade; some of us want friendship. Some of us want to join together in the pursuit of truth and understanding. Twitter gamifies discourse and, in so doing, offers us re-engineered goals for our communicative acts. Twitter invites us to shift our values along its pre-fabricated lines. We start to chase higher Likes and Retweets and Follower counts — and those are very different targets.

Others of us may come to Twitter already interested in popularity and status. For those, the gamification of Twitter may not represent such a radical change in the basic content of their goals. But even for those already interested in popularity, Twitter can change the way in which they conceive of popularity – by making highly salient a handful of specific metrics for popularity. Like and Follower counts are not the only way to conceive of popularity, but they the measure that Twitter highlights.

What's more, the effect of Twitter's gamification, across the community Twitter users, will tend towards levelling and flattening the diversity of values. Insofar as Twitter's gamification motivates its users, then it will drag all of its users' communicative values in the same direction – towards the same metric. Gamification homogenizes the value landscape. And this phenomenon will help explain some of the more socially toxic aspects of Twitter. The technology invites us to focus our cares on the narrow task of getting points and going viral. And that goal is in tension with our interest in having morally sensitive and openhearted communication. This gamification invites us, instead, to view communication through the

lens of competition, victory, and success on Twitter's very specific terms.

Let me emphasize the fact that Twitter offers us an invitation to change our values. Twitter will not change our values for us. It is a system designed to offer us pleasure in return for simplifying our values – but we still have to take up that offer. But it does ease the way for us considerably, by offering a pre-prepared and seductively designed pathway.

Of course, Twitter isn't the only place where gamification influences communication, discourse, and collective understanding. We can see similar effects with Facebook's Likes, YouTube's clickthrough and watchthrough counts,[5] academic citation rates, and more. But here I wish to explore, in details, how gamification impacts discourse and knowledge-production in one particular instantiation, as an opening step towards understanding life in the time of quantification.

Games and Gamification

Why gamify? If there is a Bible to the contemporary gamification movement, it is Jane McGonigal's book, *Reality is Broken: How Games Make Us Better and How They Can Change the World.*[6] McGonigal provides a clear — and very influential — argument for gamification. Ordinary life, she says, is quite painful. Everyday activities, like work, education, and chores, are dull and repetitive. But luckily, she says, we already have an extremely effective technology for eliminating drudgery: computer games. By importing key design features from mod-

[5] The YouTube examples were suggested by Mark Alfano.
[6] For a more technique-oriented design manual for gamification, see Chou (2015).

ern gaming into ordinary activities, we can transform daily life into something far more enjoyable (McGonigal, 2011).

In so many modern computer games, we voluntarily engage in what looks, from the outside, like pure drudgery. Many games, particularly computer role-playing games, involve what's known as "grinding" — performing simple, repetitive activities to slowly build up various in-game points and currency. Grinding can involve killing easy enemies, over and over again, for experience points and gold — or laboriously gathering piles of ingredients in order to craft equipment. Why are people willing to engage in such drudgery in their spare time when they avoid such activities like the plague in real life? The answer seems to lie in the powerful feedback and reward mechanisms available in games, especially contemporary computer games. In such games, we are given immediate rewards for our achievements in the form of points, leveling up, achievement badges, and the like. Games *quantize* our successes, making our progress clear and vivid. McGonigal emphasizes, in particular, how games offer us a steady sense of progress and victory, through a constant stream of clear feedback, in the terms of the accumulation of points (52-63).[7] So why can't we borrow those feedback and reward mechanisms, and slather them over real-life activities?

We've seen, in recent years, many efforts to gamify the workplace and the school. Business entrepreneurs seem particularly interested in gamification's ability to increase worker productivity by increasing worker motivation. Disney famously gamified its hospitality

[7] The mechanisms for this are complex. McGonigal provides a survey of the empirical literature; for a more pessimistic counterpoint, see Schull's work on game addiction and its relationship to points. Schull's account stresses the way in which the exact timing of the quantized reward in certain game designs triggers addictive surges of serotonin.

workforce, providing leaderboards and rankings for speedy performance. Notably, the system increased productivity, at the expense of also increasing the injury rate. And workers hated the system, calling it the "electronic whip" — and saying that they couldn't help being motivated by it, even though they detested the intrusion (Gabrielle, 2018). We've seen the introduction of gamification into fitness, with technologies like FitBit and Strava offering game-like structures of points, rankings, and leaderboards for exercise. We've seen gamified education in schools, and in various apps. The popular language learning app DuoLingo gamifies language learning by offering its users points and virtual medals for achieving various daily goals, like learning new vocabulary words.

McGonigal and her fellow gamification advocates are optimistic about the utopian potential of gamification. In McGonigal's picture, gamification is an unalloyed good: it simply removes drudgery and adds pleasure. But her optimism depends on believing that gamification can achieve these psychological goods while adequately preserving the value of the activity.

When we understand the source of gamification's motivational power, we will see the problem with McGonigal's optimism. Gamification involves a trade: it increases our motivation in an activity by narrowing and simplifying the target of that activity — which, in turn, changes the nature of the activity. And this may be fine when the activity has a naturally simple target, as is possibly the case with language learning. But the goals of discourse are many and subtle, and gamification threatens to destroy much of that diversity and subtlety.

The usual view among gamification advocates is to treat games and gamification as providing the same sort of value. Insofar as games are good, the story goes, then gamification must also be good, since it makes life more like a game. But this view conceals the profound

differences between games proper and the gamification of real-world activities. To understand that, we'll need a clearer account of the nature and value of games.

Let me summarize my account of games, which I have developed in excruciating detail elsewhere.[8] Games, I've argued, are the art form that works in the medium of agency. The game designer doesn't just create characters, stories, and environments. The game designer sculpts *the temporary agency that the player will occupy during the game*. They design, not only a world, but *who the player will be* in that world. I do not just mean that the game designer provides a fictional backstory for a character. They design the essential agential structure of the in-game actor. They designate what the in-game agent's abilities and affordances will be — whether they will be a jumper, a shooter, a builder or an information gatherer. And, most importantly, the game designer sets the in-game agent's *motivations* by setting the goals of the game.

And the game player submerges themselves in this sculpted agency, temporarily. Game-playing involves the temporary adoption of an alternate set of goals. Why do all this? For one thing, our goals in game-life are so much clearer than in ordinary life. In ordinary life, our goals are often obscure. We often don't know exactly what we're doing — or we find our reasons hard to articulate and difficult to apply. And we are beset with a confusing welter of values – both from within our own value system, and from the bruising value complexity of the social world. But games offer a relief from all that. While playing a game, we know exactly what we are trying to do — and afterwards, we know exactly how well we have done. Success in a game is clear and unmistakable. There are points.

And game values usually fit neatly with one another. In ordinary life, our values are hard

[8] The present account relies on material drawn from Nguyen (2017; 2018; 2020).

to balance. I care about spending time with my loved ones, raising my children right, writing good philosophy, enjoying myself in rock climbing, staying healthy, and eating delicious food. Not only are my values often in tension, but there is usually no way to precisely compare them. How do I compare achievements under one of these goals against sacrifices in another? What, exactly, is the cost-benefit analysis for choosing between working today or taking my children to the aquarium? But with games, there is usually a clear central currency of value. A game tells me to achieve victory points and then tells me exactly how many victory points things are worth.[9] The goods of a game are readily commensurable, by design.

In ordinary life, values are often inchoate, subtle, and difficult to apply. But in games, values are easy. Games offer us a momentary experience of *value clarity*. They are a balm for the existential pains of real life. In games, we know exactly what we are doing and why we are doing it. And when we are done, we know exactly how well we have done. Games offer us a momentary respite from the value confusion of the world.

It is relatively easy for the game designer to create value clarity, because the values in games are entirely artificial. The game designer can just tell us what to care about, and players simply care about it for a while. This is part of what it means to say that agency is the medium of games. The in-game agencies — their abilities, their motivations — are the plastic medium which the game artist manipulates to achieve their effects. But when we seek to gamify ordinary life, we are trying to impose value clarity on a pre-existing thicket of values. This is the worry with Twitter. Twitter can grant us the emotional security and existential

[9] Some other games offer a few different currencies of success, but even then, those various currencies are usually compatible. In many computer role-playing games, for example, I am offered both experience points and gold, with no clear explanation of which I am to pursue. Though there is no direct exchange rate between the two currencies, they go hand-in-hand. Usually the path to more experience points is through more gold, and vice versa. So for success I can aim to maximize both.

relief of value clarity, but we must adopt Twitter's narrowed targets in exchange.

How Twitter changes discourse

McGonigal views gamification as providing nothing but a motivational boost. The analysis I've offered shows the problem with that view. We get those extra motivational elements — pleasure, fun, engagement — in exchange for substantively changing the goals of the activity, and so changing the activity itself. The gamified design of Twitter influences discourse by inviting its users to *change the goals of their participation in discourse* — to simplify those goals in exchange for pleasure.[10]

Let me stipulate a bit of terminology. Let us call the designed technology which offers points and scores "design for gamification". And let us use "gamification" to refer to those cases when a player interacts with design for gamification and actually adopts those points and scores as primary motivators during the activity — when the activity actually does become something like a game for them. Notice that you can need to actually adopt these clear goals, at least for the moment, to get the pleasures on offer.

Consider some of our ordinary goals for communication. We may wish to collectively

[10] My view here is much opposed to Ian Bogost's famous argument that "gamification is bullshit". Bogost's argument here is that gamification is bullshit because the term 'gamification' was used so flexibly and variably by corporate profiteers that the term was essentially useless — that it was a pure buzzword, with no content (Bogost, 2011). As my discussion shows, gamification is a specific phenomenon with clear techniques and identifiable consequences. The fact that salespeople have used the term poorly does not undermine the usefulness of the term itself. Interestingly, I do think that gamification *is* bullshit, but in a different sense. Harry Frankfurt's discussion of bullshit can be read in the following way: bullshit is an activity that has been diverted from its usual goal (Frankfurt, 2005). In that specific sense, I do think gamification is bullshit. (Bogost cites Frankfurt's discussion of bullshit, but Bogost misses much of the specificity of Frankfurt's analysis.)

pursue truth and understanding, or to promote empathy for one another. But Twitter's scoring mechanism invites us to replace those values with another, much simpler goal: that of maximizing one's Likes, Retweets, and Follower counts. Twitter's measures are a radically simplified — and quite impoverished — rendition of the wide plurality of values for communication we might hope to find across a community of conversers. For one thing, we have evidence aplenty that what makes something go viral is not its truth, or the degree to which it promotes understanding. Recent studies have shown that tweets loaded with strong moral emotions, like outrage, are far more likely to go viral, via an effect that researchers call "moral contagion" (Brady et. al., 2017).

But a gamification booster might resist this portrayal. They might suggest that Twitter's scoring mechanism does an adequately good job of reflecting the true plurality of values. Perhaps individuals have their own values for which they come to Twitter. But those values guide when each individual user decides to Like, Retweet, or Follow. Thus Likes, Retweets, and Follower counts serve as useful measures of overall success against a plurality of values, since they function to aggregate individual approval.

But being guided by an aggregate measure of the audience's approval is a far cry from being guided by one's own internal sense of value. First, pressing Like is a quick reaction. It typically records a user's positive first-impression response to a tweet. So if we evaluated our communicative attempts by their Like counts, we would be effectively biased in favor of tweets that users immediately enjoy. We would be effectively biased against slow-burn content — against those ideas that lingered in the memory and revealed their depths slowly. It seems far more likely that a user will Like a tweet if it, say, expresses a view that they already agree with, than one that presents a challenging or subtle view that the user will have to

wrestle with for a while. This is not because tweets somehow can't be profound by their very nature. Rather, it is a feature of how Twitter's interface captures the data to feed its metrics. A user might eventually come to appreciate a challenging tweet, but they are far less likely to go back and find that tweet, weeks later, to press Like. Slow appreciation is far less likely to be captured by the system and be counted towards that tweet's score. And insofar as we have become gamified, then we will judge our own communicative success in terms of that recorded score.

Second, Twitter scoring emphasizes the total number of Likes, rather than, say, the depth of engagement or lasting effect of a particular communication. This sort of problem plagues all sorts of large-scale value aggregations. Consider Matt Strohl's criticism of the movie-review aggregator site Rotten Tomatoes. Rotten Tomatoes surveys the online reviews of a movie and reduces them each to a simple binary: was it a positive or negative review? And then Rotten Tomatoes produces an aggregate percentage of positive reviews. Notice, says Strohl, how this influences the results. A movie which strikes every critic as a little bit above average will score 100% on Rotten Tomatoes and show up at the top of the heap. A movie which divides the critics — which some critics find utterly brilliant and other critics find baffling — will show up with a 50% score, and appear, numerically at least, as a mediocrity. But great movies, says Strohl, rarely please everybody. Much of the most important art is difficult and utterly divisive. But the filtering and aggregating mechanism of Rotten Tomatoes ends up expressing a mathematical preference for more blandly agreeable material (Strohl, 2017).

Twitter's aggregation method produces a similar effect. Sometimes, when I'm teaching, I

say something to a whole class that I doubt will reach most students, but that I strongly suspect will resonate with one or two students. And often, that's good enough for me. But Twitter scores each tweet with a simple binary measurement: either we Like a tweet, or we don't; either we Retweet, or we don't. This binary data collection screens off, at the input stage, any considerations of depth of impact or profundity of connection. Then Twitter automatically, and very visibly, aggregates the results of that binary input. Twitter's scores make highly salient the *number* of users with positive reactions, while de-emphasizing the *quality* of any particular interaction. Insofar as we come to be motivated by Twitter's scores, then the aim of our communication will be subject to a similar biasing effect as with Rotten Tomatoes. We will prefer those communications that appeal to the greatest number — even if that appeal is marginally positive — rather than those communications that might reach a smaller number more deeply.

Third, Twitter scoring aggregates user interests into a single monolithic statistic, which threatens to diminish the plurality of values for which we collectively communicate. Let's assume, for the moment, that every Twitter user Likes those tweets which in a way that accurately reflects their particular interests in communication. (In other words, assume that the last two problems don't apply.) Even so, gamification will result in a homogenization of the values for which various actors communicate. Pre-gamification, each tweeting user will be motivated by their own particular values in communicating, giving us a diversity of communicators with different and distinctive motives. Such a diversity of interests is quite healthy, epistemically speaking. Cognitively diverse communities do better at figuring things

out.[11] Suppose, now, that the entire community succumbs to gamification and start chasing popularity by Twitter's metric. Post-gamification, we have a body of communicators identically motivated to satisfy the same mixed populace. We've replaced a diversity of motivations with a motivational monolith. (Here's one way to make the damage apparent: imagine the difference between a world of artists each motivated by their own aesthetic sensibility, versus a world of artists each motivated to satisfy the largest number of their fellow artists.[12])

These three arguments approach, from different angles, the same central idea: that Twitter's scoring mechanisms offer a simplified rendition of the rich plurality of our values. They refract our interests through the particular prism of Twitter's information collection system, and then average the result. And, insofar as this simplification comes in an attempt to re-optimize the activity for pleasure, we should expect it reduce that activity's capacities to perform its other functions. At most, we might have hoped for a compromise between pleasure and the original goals of the activity; but such a compromise would require a careful, intentional design effort. And we have little evidence that Twitter's design for gamification arose from any such careful attempt to support the plurality of communicative values. We have, instead, plenty of reason to think that its design features were heavily driven by an interest in increasing user engagement for the sake of profit.

Here's an analogy. Products that seem good and exciting in the store often turn out to be quite poor in quality. This superficiality is no accident; it is the result of systematic pressures. The function of these objects has drifted due to market forces. Where once the function of a

[11] Lu Hong and Scott Page (2004, 2007) have famously demonstrated that cognitive diversity trumps cognitive ability in groups of deliberating individuals. For a rich application of these results to political communities, see Hélène Landemore's (2013, 89-117) discussion of inclusive deliberation.

[12] For a further discussion of the relationship between loyalty to personal aesthetic sensibility and a resulting landscape of creative diversity, see Nguyen (forthcoming).

shoe was to help us walk, now, so often, the function of the shoe is to get bought.[13] And what gets a shoe bought is not its actual long-term quality, but the short-term appearance of quality. I am suggesting that something similar happens with gamified discourse on Twitter. Gamification changes discourse from serving the long-term values of communication to serving the function of gathering the most Likes and Retweets.

Of course, gamification might not be dangerous if it is managed properly. Here, then, is a more sophisticated defense for the gamification optimist. Perhaps the simplified, gamified value isn't actually *replacing* our original value, but simply functioning as a short-term *heuristic* for that value. Cognitively limited beings like us often need to focus on a short-term proxy for a complex value – like, say, using one's increased running mileage as a proxy for health, or using one's grades as a proxy for educational success. Managed properly, such heuristics can serve as an efficient and motivating proxy for some deeper and more complex value. It's much easier, on a day-to-day basis, to aim at increasing mileage than it is to think about my health as a whole.

But proper management is key. Heuristics, after all, are simplifications of the real thing. They are good heuristics insofar they remain properly tethered to our deeper values. The successful use of a heuristic involves a complex process of management. We need to step back and reflect on whether using the heuristic is actually helping to achieve the underlying values. Increasing your running mileage might sometimes be a good proxy for fitness, but not when it brings irreversible knee damage. We need to adjust our heuristics when they drift.

But Twitter's design for gamification discourage appropriate management. First, Twitter

[13] This excellent formulation suggested by Alison Rieheld.

makes the scoring system pervasive and highly salient through its user interface. The ready availability of this pre-fabricated, neatly packaged evaluation system may, by itself, discourage further reflection on one's values. More importantly, Twitter's metric is hard-wired into the system. Even if we managed to discover that we did, in fact, want to adjust the heuristic, that adjustment is hard to do on our own – because the scoring system is embedded in an externally-controlled technology. Gamification works on us, in part, because of the ready availability of those quantified evaluations.[14] In order to get the full pleasure on offer, we must assent to the particular measures that have been baked into the system. So, unless Twitter's gamified metrics just happen to the right heuristics to achieve our particular values in communication – and to do so in perpetuity – then taking on those metrics will impede our capacity to manage our proxy targets in light of our real values. And everybody who uses Twitter is pressured to take on precisely the same heuristic, with little room for personal tailoring. Twitter's interface comes with a pre-fabricated, hard-wired measure, which points its users firmly in the direction of popularity – rather than allowing the user to search out the heuristic that best matches their own interests. Of course, a user could, conceivably, resist the pull of Twitter's design for gamification and impose their own self-created and self-managed heuristics on their tweeting. Twitter isn't actually forcing a value change on us. The system design is seductive, but not compulsory. But Twitter offers us an intoxicating hedonic reward for changing our values along its pre-arranged lines.

These thoughts about pre-fabrication points the way to another worry. To provide the kind of carefully engineered, automated, steady feedback that McGonigal praises, we usually

[14] See McGonigal's (2011, especially 52-63) discussion of *World of Warcraft*, and the importance of the visible and steady trickle of points and rewards.

need help from large institutions – like corporations and governments. Our choice of activities – and the way we engage in those activities – will then depend deeply on the fabrication efforts of large-scale institutions. Suppose that you find yourself with a strong preference for gamified activities over ungamified ones. When you engage in a particular gamified activity, you will, indeed, find yourself more motivated and more engaged. But you will also be restricted to choosing from the list of activities that institutions have chosen to gamify for you. Right now, for example, there are popular gamifications available for language learning, increasing your step counts, and tracking your weight loss. But there aren't good gamifications for learning to appreciate complex poetry or becoming a better and more empathetic listener. And even if we think that the choice of gamifications isn't insidious or manipulative – even if the institutions are well-intentioned and just trying to help us lead our best lives – we will still find that the range of activities available to us will be sharply curtailed. Institutions will tend to produce gamifications of activities that more easily admit of technologized measurement. It is easier to gamify weight-loss than it is to gamify deep aesthetic appreciation, because the former is easier to measure in an automated way. And when institutions gamify activities with more subtle and complex aims – like communication – then they will tend to tend to change those activities to make the aims more amenable to automated measurement. So a life of gamification will tend to draw us towards those activities which have clearly measurable goals, or can be transformed into something with clearly measurable goals. When we demand the pleasures of gamification in our activities, then the range of activities available to us diminishes – and the degrees of freedom we have within the activity also diminishes. Ironically, if we took the spirit of *play* to involve something like some kind of freedom or spontaneity with respect to one's values and activities, then gamification turns out

to be the opposite of play.[15]

How gamification changes us

So far, we've discussed how gamification can change the goals of the activity and so change how we conduct the activity. There is now a further question: how might gamification change the users themselves? How might they transform the users' lasting values?

Much depends here on how the users motivationally interact with the scores. There are several different ways that interaction could go. First, users can treat Twitter as a game proper, taking on its goals temporarily for the sake of the pleasure during the activity. Second, they can internalize those scores and transform their long-term goals for communication. Third, they could keep the scores at motivational arm's length, treating them only as a measure of some useful resource, but not permitting the scores to function directly in their motivation in any way. Let's look at these various possibilities one by one.

First, suppose one treats Twitter as a game proper. Let's call such a person a *game-playing user*. Such a user temporarily adopts Twitter's scores as their goal while they play Twitter, and then puts those goals away afterwards. In that case, their local adoption of those game-goals wouldn't count as a long-term change in their value. And this practice would be perfectly harmless, if Twitter were really a game through and through — but Twitter is not.

Real games have special properties. As Johann Huizinga famously put it, a game occurs in a separated place — a place he called "the magic circle" — where we take on alternate roles, and our actions took on alternate meanings (Huizinga, 1971). Or, as Annika Waern puts it, games take place inside an interpretive frame, where we agree to reinterpret the meanings

[15] This view of play is fairly common in the literature, but this precise articulation is from Maria Lugones (1987).

of the acts inside the game (Waern, 2012). These are philosophically rich descriptions of a familiar phenomenon. Actions in games are screened off, in important ways, from ordinary life. When we are playing basketball, and you block my pass, I do not take this to be a sign of your long-term hostility towards me. When we are playing at having an insult contest, we don't take each other's speech to be indicative of our actual attitudes or beliefs about the world.[16]

And there are, in fact, conversational practices that are games, through and through. These are explicit, temporary practices where we conduct conversation while taking on specific goals, obeying specific obstructions, and taking on specific roles. There are structured games of deceit, intended to be played at parties or as tabletop games, like Mafia, Werewolf, *The Resistance: Avalon,* and *Spyfall*. There are also informal conversational games, like when we sit around and try to come up with the best insult about each other's mothers. What makes the deceit in true games morally permissible is that we all know, going in, not to take the in-game speech seriously.[17] I don't actually take your "Yo mama" insults to be presented as reliable testimony about the state of the world. Such games involve the voluntary and consensual entrance, by all the players, into an alternative game-space, where the players know to interpret the actions and communications inside the game under a special light – to not treat them as ordinary, real-world actions.

[16] The "magic circle" notion has come under significant fire, which I believe can be located in a famously overstated version presented in *Rules of Play*, an influential early game-design textbook. According to that textbook — at least, according to some readers — magic circles were *impermeable* membranes for meaning, across which no moral judgment or consequence could cross (Salen and Zimmerman, 95-97). I am relying here on what I take to be a more minimal and defensible version of the magic circle. For various defenses of more reasonable accountings of the magic circle, see (Stenros, 2012; Waern, 2012; Nguyen 2020, 177-180).

[17] For further discussion of the moral transition into game-life, see (Weimer, 2012; Kretchmar, 2012; Nguyen, 2017).

But those aren't the normative conventions around most of Twitter. The majority of Twitter presents itself as, and is taken to be, ordinary discourse. For the most part, we think that people on Twitter are representing their real beliefs and trying to make claims about the actual world.[18] A user who approaches Twitter as a literal game, then, runs the risk of undermining the epistemic goods available to the other users. Suppose I'm on Twitter to actually communicate about ideas, and you're playing a game with Twitter — saying whatever it takes to get the most Likes and Retweets for the sheer fun of it. If I don't realize you're playing a game, then I will be profoundly misinformed by your tweets. Those who approach Twitter explicitly as a game, but don't clearly mark themselves as game-players, are conversing in bad faith. They are presenting themselves as engaging in a discursive, epistemic practice while actually being guided by non-epistemic motives. And, insofar as other Twitter users take game-playing users as serious participants in sincere discourse, then these others will be mistaking gaming talk for serious testimony.[19]

Second, users could internalize the scores of Twitter, permitting their enduring goals with to be influenced by Twitter's scoring mechanism. Twitter makes this easy, by making those scores so prominent and so pervasive.

---

[18] An exception is so-called Weird Twitter, which is a sub-network devoted to irony and verbal game-playing, largely its own, largely segregated network. But that is, of course, a specific exception — and not a particularly risky one, since Weird Twitter tweets are so bizarre and incomprehensible, that they are not likely to be mistaken for ordinary discourse. Weird Twitter is a game proper, with clear indicators that the users as just playing around. But things are different elsewhere Twitter. I suspect that many other people are playing a game with Twitter with political speech, on the main stage of Twitter, which is a far more dangerous affair.

[19] To put this into the technical language of the epistemology of testimony: I take Twitter to be a context in which most tweets are treated as "assertions". To use Elizabeth Fricker's account: when S asserts that P to an audience H, they thereby vouch for the truth of P to H, presenting P as being so, such that H can form belief that P on S's say-so. An assertion that P represents the asserter as knowing P (Fricker, 2006). Game-playing users, then, are presenting non-assertions into an assertoric context, where others can be reasonably expected to treat them as assertions.

Twitter is a part of a larger phenomenon here, which we can call *value capture.*[20] Value capture occurs when:

1. Our natural values are rich, subtle, and hard-to-express.

2. We are placed in a social or institutional setting which presents simplified, typically quantified, versions of our values back to ourselves.

3. The simplified versions take over in our motivation and deliberation.

Some examples: starting to exercise for the sake of your health, then getting captured by FitBit and coming to just care about your daily step-counts. Going to school for the sake of a good education and coming out obsessed your GPA. Becoming a pre-law for the sake of public interest and legal activism, and then coming to care more about getting admitted to the best law school according the *US News & World Report's* law school rankings.[21] And, of course, going onto Twitter for the sake of communication, connection, and shared understanding — and coming out obsessed with maximizing Likes, Retweets, and Follower counts. And, obviously, a high step-count isn't the same as good health; a high GPA isn't the same as a good education; and high Twitter Likes aren't the same as connection or collective understanding.

Value capture occurs when our values undergo a long-term and enduring simplification, as guided by the external metrics provided by institutions and technologies. The worry here isn't that our values couldn't ever be expressed in quantified form, in principle. Rather, it's

[20] I introduced the notion of value capture in (Nguyen, 2020, 189-215), though my views have, I think, matured since that earlier sketch.

[21] Wendy Espeland and Michael Sauder's *Engines of Anxiety* is a thorough - and deeply alarming - account of how the *USN&WR*'s law school rankings have profoundly changed student motivations (Espeland and Sauder, 2016).

that the kind of metrics, measures, and gamified scoring that we typically encounter in our life with bureaucracies, institutions, and corporations are almost always radical simplifications of the values they claim to be measuring. Those simplifications may have certain uses in administration, management, or large-scale scientific data-collection. But what makes them useful for those functions is, in fact, their very simplification.

It's useful here to borrow from a nearby discussion: that of the simplifications involved in bureaucratic quantifications. As Theodore Porter puts it, institutional quantification is driven by an interest in making information highly portable. Rich, nuanced qualitative information is difficult to manage from any sort of informational center. We need to strip out the context-sensitive details and nuance in order to transmit it easily between contexts (Porter, 1996). This is why such quantification is beloved of centralized bureaucracies, which need to pass information to distant managers, and up many levels in the hierarchy of administration (Scott, 1998).[22]

This context-stripping standardization also allows us to aggregate the information arithmetically. Think, for example, of how teachers assess students. Teachers could offer each student rich and individualized commentary about the strengths and weaknesses of their academic work. Such individualized commentary would be vastly useful to the students. But such individualized commentary is incredibly hard to aggregate and manage by upper level administrators. How is an administrator supposed to compare the portfolio evaluations of the art department with the mathematical performance of the statistics students? So teachers are asked to provide quantified grades for their students, which can then easily be averaged across classes for one students, and across all the students in a department, university,

[22] I am also influenced here by (Perrow, 2014; Merry, 2016).

or school district. Quantified grades strip out much of the most important information. But that context-stripping renders information into a standardized form that can be operated upon arithmetically. This allows managers of massive, sprawling institutions to bring their entire domain into view — by putting information in standardized form and then aggregating it. Institutional life exerts a pressure on information, pushing it towards quantified, aggregable form. Notice that these forces do not typically arise to support our individual interests, but instead the interests of management and large-scale administration. But, problematically, those quantifications appeal to our motivations precisely because of their apparent clarity. And once we offer simple, quantified metrics for success, those metrics take over in the motivations of so many people.[23]

A similar pressure occurs with overt gamifications, especially ones in an automated, technological context — though the motivations for simplifying may be slightly different. In order to create the motivational rewards of gamification, we need to provide a score. In order to provide that score, we need to offer reliable scoring mechanism. And in a large-scale, technologized context like Twitter, that scoring mechanism needs to function automatically. Twitter can't offer a score based on quality of engagement, empathy, or depth of thought. It can only score us on what is easily legible to its systems: like whether or not somebody clicks on Like.

Such scores present a game-like motivational lure. But because they can also be inte-

[23] For excellent case studies into the motivational pull of quantifications, see Sally Engle Merry's (2016) study of the use of simplified metrics and indicators in motivating political action; and Espeland and Sauder's (2016) study of the effects of law school rankings on the motivations of students and donors. This brief sketch of value capture, quantified bureaucracy, and seductive clarity here will be developed in future work.

grated arithmetically, they can be used to generate more gamifications down the line. Consider, for example, how FitBit's scores nest. FitBit provides me with a daily score for my walking. But that score can be averaged, so I can also get a score for my walking success each week, and each month. Since other people's scores can also be averaged, the system can automatically generate rankings and leaderboards, each of which provides another game-like motivational boost. Similarly, Twitter's scores, once rendered into quantified form, can be extracted and used to generate other scores.[24] For example, the explicitly gamified social network Empire.kred creates a second-order game out of social media scores. Empire.kred is a virtual stock market, where individuals are the stocks. Individuals can invest in each other using the game's virtual currency, $Eaves. Their stock value is based on their social media power, as modified through investments. Empire.kred harvests various scores from other social media networks — like Twitter and Facebook, and then aggregates those scores to drive its virtual stock market.[25] This is possible because those social media networks have already digested evaluations of user success into a standardized and portable format: a numerical score.

The problem with value capture can be put most clearly if we help ourselves to an assumption. There is, it is often thought, a natural aim to belief. Belief aims at the truth.[26] We can be tempted by other motivations to abandon that aim: to believe what will feel pleasant or make things easy for us. But to do so is to abandon the natural aim of belief; it is to subvert

[24] Lupton and Smith's (2017) recent study of quantified self-tracking show that many self-trackers are extremely interested in the exportability of self-tracking data — of the ability to send FitBit step-counts to a spreadsheet or more macroscopic health-tracking program. Their approach, however, is more tuned to the data-gathering side, and pays less attention to the motivational possibilities of gamification.

[25] https://play.empire.kred

[26] This idea has its most influential statement with Williams (1970). For more recent discussion, see (Velleman, 2000; Wedgewood, 2002).

the activity of believing. The activity of earnest discourse also seems to have a natural aim, which is the collective pursuit of truth. We aim to express what we think of as true, and to question and challenge each other's expressions, as part of our quest to understand the world. But gamification tempts us to change our goals — to aim at expressions which maximize our score, rather than those which aid our collective understanding. And it promises to reward us for that change with pleasure. Twitter tempts us to subvert the activity of earnest conversation for hedonistic reasons.

Besides game-playing users and value-captured users, there is a third possibility: that users could treat the scores of Twitter as simple reports of some instrumental resource, useful for the pursuit of further ends. They treat Twitter's numbers, not as setting a goal, but merely as useful data. Let's call such person a *value-independent user*. Such a user has avoided internalizing the scores of Twitter in any way. They have avoided gamification.

Here's what that might look like. Suppose one wanted public influence. A resource for public influence is having a Twitter account with a wide number of followers — and tweets which are heavily retweeted will reach a large number of people. So one could aim for high scores simply as an approximate measure of that instrumental resource. Such a value-independent user wouldn't have any form of change of value or goals, either short-term, or long-term. They also wouldn't be subject to the motivational boosts that arise from more fully inhabiting the scores of Twitter. They would be holding those scores at phenomenal arm's length. Such a user, then, would be free of the more pernicious effects of value capture and game-playing. They have resisted Twitter's invitation to gamification.

Thinking about the value-independent user helps us get clearer on what's wrong with the

value-captured user. The value-independent user manages the scores, where the value-captured user is driven by the scores. Consider, by way of analogy, two relationships you could have with money. First, you could view it as an instrumental resource, to be collected in pursuit of some other value. Second, you could treat it as an enduring end, to be pursued for its own sake. Somebody who sought money as an instrumental resource would manage their pursuit of money in view of their larger ends. Somebody who pursued money as an instrumental resource to happiness, wouldn't take that high-paying job that would destroy their happiness. They would manage their pursuit of money, making sure to pursue money only to the extent that it actually helped their happiness. The person who pursues money for its own sake, however, has no such guiding purpose with which to manage their pursuit of the greatest pile.[27]

Similarly, consider a user who comes to Twitter for the sake of, say, social progress, and sought Followers and Retweets simply as an instrumental resource for their mission. They have an external standpoint from which to manage their pursuit of Followers and Retweets. They wouldn't say *anything* in order to go viral, for many such things they could say would likely undermine their larger purpose. But the person who has been fully value-captured by Twitter's scores has no such limitation. They will be driven to say whatever it takes to go viral and get those points.

For the value-independent user, Twitter's scores are merely a means. But for the value-captured user, Twitter's scores have become the end. The act of communication itself has been instrumentalized to the end of Twitter scores. Rather than using Twitter scores to advance their independent values in communication, they have changed the nature of their

[27] This paragraph arises from discussions with Aaron James.

communication to advance their pursuit of Twitter scores.

Changing values

The key notion here is the idea that gamification problematically *instrumentalizes* our goals. The notion of instrumentalization will be useful understanding some of the socially toxic behavior which seems to bloom on Twitter. But first, we'll need a clearer picture of the notion of instrumentalization. For that, we'll need to look, in greater depth, at how and why we fashion new goals for ourselves in games and gamification.

My account of games shows that the player has a rather extraordinary form of agential fluidity. During a game, a player takes on an alternate agency with alternate goals. That agency has been engineered to provide satisfaction for the player who adopts it. Gamification works in a similar way — it offers us various satisfactions, in exchange for shifting our goals along its engineered lines. Both games and gamification involve instrumentalizing our goals. This is unproblematic in games, but deeply problematic in gamification. Why? Because games are a very peculiar and distinctive sort of activity, and gamification doesn't share in some of the most important features.

The best account of the special nature of games comes from Bernard Suits' marvelous attempt to define 'game'. Suits says that to play a game is to voluntarily take on obstacles to make possible the activity of struggling to overcome them. In other words, in a game, the obstacles are much of the point. We try to run a marathon, and what it is to run a marathon is to try to get to a certain place while submitting to various restrictions. We must run by our own power only – no short-cuts, no taxis. Those restrictions help constitute various obstacles

for our efforts. But, says Suits, our devotion to these restrictions shows that we are not motivated simply by the independent value of crossing the finish line. If we just cared about being at that particular point in space, in and of itself, we would take the most efficient means to that end — like a taxi. The fact that we are willing to place extra, unnecessary inefficiencies in our way indicates that our interest is not in actually achieving the goal in and of itself, but in achieving it inside certain specified restrictions. Our interest is to achieve the goal *by way of a particular, constructed form of activity.* As Suits puts it, in a game, the restrictions help constitute the very activity we are interested in performing. What it is to run a marathon, is to run a certain distance under one's own power. If we took a taxi, we wouldn't be running a marathon at all (Suits, 2014).[28]

As I have argued elsewhere, Suits' account reveals the possibility of a very peculiar motivational structure.[29] There are two different motivational structures for playing a game. One could be an *achievement* player, who plays the game for the value of winning. Or one could be a *striving* player, who takes on a temporary interest in winning for the sake of engagement in a struggle. (One could also play for both motivations, in varying proportions.) Striving play is a very special motivational structure; it involves a motivational inversion from ordinary life. In ordinary life, we take the means for the sake of achieving the ends. But in game life, we select the ends for the sake of the means. We take on a temporary end, and we submerge

[28] I don't think gamified activities count as games proper for reasons that are tangential to the topics for this paper. Briefly, according to Suits' definition, which I largely endorse, games are activities where the *goal* of the game is partially constituted by the *designated restrictions* on that goal. What it is to make a basket in basketball is, in part, constituted by the player's having obeyed the dribbling restriction. For a further discussion of this point, see Nguyen (2020, 27-73). The goals in gamified activities are not restriction-constituted in this way.

[29] This is a very brief presentation of one argument for the existence of striving play, among several I have offered elsewhere. The most detailed version of this analysis occurs in (Nguyen, 2020, 27-73).

ourselves in it.[30]

When my spouse and I play games, we want to both have a good time, so we look for games that we're both relatively good at. We can see the fact that we're both striving players by how we manipulate our capacity to win in the long-term. Suppose that we have found a game at which we are perfectly matched and are having a lovely set of intense gaming sessions with. Suppose one of us finds a strategy guide to that game. If that person were to read it by themselves, they would pull ahead and start winning. If we were achievement players, then we each should want to read that guide. But we don't, and it is perfectly reasonable that we don't. We are willing to suppress our capacity to win in the long-term — even though we try, with all our might, to win during the game. Our extra-game behavior reveals that we aren't actually interested in winning in any enduring sense. Our interest in winning is merely something we temporarily adopt, in order to create the experience of that delicious struggle.

And the goal we pursue in the game is often disconnected from our enduring goals and ends — at least, disconnected in the usual linear sense. In many games, our real purpose is to have fun, but we can only have fun by trying to win. But we don't really care about winning; we just adopt a temporary interest in winning so that we can engage in the fun activity of trying. But after the game is through, we can dispense with that interest in winning. For example: I can start a game of Charades at a party for fun. In order to have fun, I have to genuinely try to achieve the goals of the game — to communicate concepts through gestures, without speech. But after the game, I discard that desire. After all, if I lost at Charades, but we all had a good time together, then I achieved my true purpose. Only an especially poor

[30] Some philosophers may protest that I have posited the impossible: that we can desire at will. Please see Nguyen (2019, 451-455) for my argument that striving play reveals that we can, in fact, desire at will.

sport would think the whole enterprise a failure because they had lost at Charades.

So, when we justify our game goals in striving play, we do not do so in reference to the value of the goal itself, or to what follows from it. We justify the game's goals by pointing to the value of the activity of pursuing those goals. Thus, striving play *instrumentalizes* our adoption of goals. In striving play, we adopt a goal, not for its own value. Our adoption of a game-goal is justified in terms of the activity of pursuit that goal structures.

Here, then, is a key difference between games proper and the gamification of non-game life. In striving games, the goals of games are temporary. More importantly, they are disconnected from the network of our enduring ends. In striving play, my in-game goal is winning, but I don't actually care about winning in the long-term. I achieve my real purpose – fun, satisfaction, exercise – by pursuing the win, and not by actually winning. And this is why it is perfectly permissible for game designers to change the goals of game-activity. Game-goals can be made as simple and narrowed as is convenient because they aren't directly attached to our enduring ends. Game designers are changing the play-goals that guide an artificial activity, which has been screened off from many real-world consequences.

But gamification is an entirely different matter. In gamification, the designers are instrumentalizing the goals of our real-life activities. FitBit, by gamifying exercise, invites us to change our goals for our health and fitness. And Twitter, by gamifying discourse, invites us to change our goals for conversation, communication, and declaration. Instrumentalizing one's goals is fine in striving games, because the goals in games were never valuable, in and of themselves, in the first place. But in real life activity, the goals are often independently valuable. So when we gamify those activities and instrumentalize those ends for the sake of pleasure, we risk losing sight of the real importance of the activity. Twitter's gamification

changes our communicative goals away from understanding, connection, and the collective pursuit of truth, and bends them towards something much more impoverished.

Twitter and toxicity

I've discussed elsewhere two problematic social phenomena associated with polarized discourse: echo chambers and moral outrage porn. Both of these phenomena seem to flourish on social media. We are now in the position to offer the beginnings of an explanation for this relationship. Gamification, echo chambers, and moral outrage porn all share a common central thread: a willingness to instrumentalize what ought not be instrumentalized.

Let's first get clearer on these other phenomena. First: as I've argued elsewhere, echo chambers are best understood as structures of manipulated trust. Echo chamber members have been systematically taught to distrust everybody on the outside (Nguyen, 2018).

To put it more formally: an echo chamber is a social structure in which:

1. One must subscribe to a certain belief system to be a member.

2. That belief system includes the belief that all non-members are untrustworthy, and all members trustworthy.

Thus, echo chambers inculcate a radical trust disparity between members and non-members. The belief system includes some explanation for why everybody on the outside is untrustworthy. In the modern landscape, those explanations often take the form of conspiracy theories — like, "The liberal media is in the grip of George Soros and totally corrupt." And

the trust disparity is self-reinforcing. The more you trust your fellow echo chamber members, the more their agreement will confirm your shared belief system. And the more you confirm that belief system, the more you will trust your fellow members and distrust outsiders.

Compare echo chambers to a nearby phenomenon: that of epistemic bubbles. An epistemic bubble is a social structure where insiders aren't exposed to views on the outside. Despite the superficial similarity, epistemic bubbles and echo chambers work through entirely different mechanisms. In an echo chamber, inside members may have plenty of exposure to outside views, but outside voices have been undermined. Epistemic bubbles are structures of bad connectivity; echo chambers are structures of manipulated credence. In an epistemic bubble, outside voices aren't heard; in an echo chamber, outside voices have been systematically discredited.

Importantly, I've argued, many problematic belief communities have been misdiagnosed as epistemic bubbles. But actually, they are mostly the result of echo chambers. It isn't that climate change deniers, for example, are simply unaware of what climate change scientist think, or the standard publicly available arguments for climate change. They are, for the most part, quite well acquainted with those arguments and conclusions. It is that they think that the institutions of climate change science have been systematically corrupted and are untrustworthy. This helps explains the intractability of climate change denialists. Since an epistemic bubble works through simply *omitting* outside voices, we should be able to shatter one simply by exposing an insider to more voices and more viewpoints. We should expect epistemic bubbles to go down with the first contact to the missing evidence. But echo cham-

ber members are pre-prepared for encounters with external viewpoints and armed with explanatory mechanisms to dismiss those other voices. Echo chambers are far more robust.

Why might one enter into an echo chamber? In my earlier discussion, I focused on the possibility that one might be raised in an echo chamber, and, through no fault of one's own, been trapped in an errant system of trust. But here I would like to focus on another possibility: that some people choose to enter echo chambers because being in an echo chamber is more comfortable and more pleasurable.

Life outside of an echo chambers is full of all kinds of cognitive difficulties. We must constantly struggle with conflicting evidence and unexplained phenomena. And we are confronted, over and over again, with evidence of our own cognitive fallibility. These confrontations humble us — which is good for us, but also quite painful.

Echo chambers banish all that epistemic friction.[31] They remove, through distrust, the impact of disagreeing voices. Instead of having to cope with new evidence, echo chambers typically present their members with clear, coherent stories about the world. Instead of the humbling confrontation with the evidence of one's errors, echo chambers offer their members the joys of unanimity and uninterrupted confidence.

And notice: these joys are very much akin to the joys of value clarity that we found in games. And both forms of joy emerge from similar engineered conditions. Games involve redesigning the agent's goals and abilities for pleasure. Echo chambers involve re-engineering their members' belief system and trust settings for pleasure. And echo chambers are dangerous because they re-engineer, not some temporary and segregated belief system, but real-

[31] I am influenced here by Jose Medina's (2012) account of epistemic resistance, though I emphasize the idea that the experience of and the processing of epistemic resistance is comfortable, and synthetic epistemic environments engineered to be resistance-less are quite pleasurable.

life belief systems which govern real-life action.

We can now see the higher-level similarity between gamification and echo chambers. In gamification, we instrumentalize our real-life goals. In particular, the gamification of Twitter involves instrumentalizing the goals associated with discourse. Gamification involves, to a significant degree, abandoning the aim of truth and understanding, and taking on a simpler goal — where that goal was engineered for the sake of pleasures of value clarity. Echo chambers also involve instrumentalizing our belief systems, abandoning the aims of having the beliefs that are true, and trusting the people that are reliable. And, in exchange for abandoning these epistemic aims, echo chambers offer their members the pleasures of confidence, simple coherence, and unity.

There is an interesting complexity in the instrumentalization here. There are two levels of explanation for these simplifying re-designs. It seems plausible, for both Twitter and for many real-world echo chambers, that they are intentionally designed by an external agent. The re-engineering involves instrumentalization at two different levels: at the level of design, and at the level of adoption. Plausibly, Twitter's makers consciously designed it for pleasure and addictiveness, for the sake of profit. So there are two instrumentalizations here. First, Twitter's makers are designing for gamification for the sake of profit, which they pursue by making their design seductively pleasurable to its end-users. And second, those users are accepting the seduction, and gamifying their discourse for the sake of pleasure. At both levels, we find people willing to forsake the original goals of discourse for some other end.

Similarly, many echo chambers are plausibly designed for political control.[32] To that end,

[32] This view might strike some as cynical. This is, however, the picture offered by Kathleen Hall Jamieson and Joseph Cappella (2010) in their meticulously researched account of Rush Limbuagh and Fox News' inten-

designers have a reason to engineer their belief system to be as pleasurable as possible. Once again, there are two instrumentalizations: designers create a belief system for the sake of political control, which involves designing them to be pleasurable to their users. Then users accept those belief systems for the sake of that engineered pleasure. And, once again, at both levels, we find people willing to create or adopt belief systems for reasons that bear, not on their relationship to truth, but to some other end.

Let's turn now to the second toxic phenomenon: moral outrage porn. In earlier work, Bekka Williams and I offer an account of "porn" in the generic sense. We mean to describe the new, modern usage, which includes things like "food porn", "real estate porn", and "closet porn". We propose that a representation is used as generic porn when it is engaged with for the sake of a gratifying reaction, freed from the usual costs and consequences of engaging with the represented content. For example: food porn is pictures of food which people look at to get immediate gratification, while avoiding the calories, cost, and hassle of eating the depicted food. Real estate porn is pictures of expensive, well-maintained homes, which people look at for immediate gratification, while avoiding the costs and hassle of buying and maintaining those actual homes.

This account helps us get a grip on an important phenomenon: moral outrage porn. Moral outrage porn is representations of moral outrage, which people engage with for the immediate gratifications of feelings of moral outrage — for the pleasures of feeling smug, secure, and confident in the total wrongness of the other side. And they do so while avoiding the

tional construction of an echo chamber. Their book, *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment,* is one of the best early analyses of echo chamber structures, and is the source for my own account of echo chambers.

costs and consequences of genuine moral engagement: like the pains of struggling to be morally sensitive, the efforts of seeking the right moral beliefs, and the exhaustion of real moral action. We think it quite clear that social media is suffused with moral outrage porn. And moral outrage porn is quite dangerous. If one is interested in using moral outrage porn for pleasure, one will have an incentive to adopt, not the right moral system, but the one that is easiest to crank for pleasure. One will likely be tempted to, say, adopt a simple and absolute moral system, that will give one the easiest access to the pleasures of smug condemnation.

Crucially, our claim isn't that *moral outrage* is bad. Real moral outrage is crucial. Moral outrage, when it emerges from a well-tuned moral sensibility, helps us to register injustice and motivates us to end it. The very problem is that moral outrage porn threatens to corrupt the real thing. The proper target of moral outrage is the genuinely outrageous. But when we use moral outrage porn, we use our own moral outrage for pleasure. And so we are incentivized to change our moral belief system — to ignore the truth, and adopt those beliefs that will give us the most pleasurable outrage. Moral outrage porn invites us to *instrumentalize our moral beliefs* (Nguyen and Williams, 2020).

So: moral outrage porn and echo chambers often occur together, and they both seem to flourish on social media. Why might that be? We now have the beginnings of an explanation. All of these phenomena involve hedonistic instrumentalization, where we take an attitude or mental state and modify it away from its appropriate target in exchange for pleasure.

Why might a similarity of motivational structure lead to frequent co-occurrence? I suggest that these various hedonistic instrumentalizations occur together because they appeal to the same sorts of motives. In other words, the co-occurrence of gamification, echo chambers, and moral outrage porn are not best explained by features of the individual phenomena

themselves, but in terms of the character of their likely adopters. In all of these cases, maintaining the attitudes towards their appropriate aim takes work. Somebody willing to abandon an attitude's appropriate aim and instrumentalize it for pleasure in one place, is likely to do it another.

Another way to put it: gamification, echo chambers, and moral outrage porn go together like junk food. Different kinds of junk food are unhealthy in different ways — some are too high in salt, some too high in fat, some too high in sugar. But the reason they are often consumed together is that they are all likely to be consumed by somebody who is willing to trade off health and nutrition in return for a certain kind of quick pleasure. The same is true of gamification, moral outrage porn, and echo chambers. They are all readily available sources of a certain quick and easy pleasure, available to anybody willing to relax with their moral and epistemic standards.

Next, think about things from the point of view of the system designer. Imagine yourself into the shoes of a hostile manipulator. Let's say you wanted to get people under your political sway. You'd want to design a belief system that was as maximally catchy and sticky as possible. Here's one way you could do it. First, you could design a belief system that included provisions to distrust all outsiders who didn't share the belief system. You could make that belief system utterly clear and coherent, all the better to please its adopters. In other words, you'd design an echo chamber. Second, you could rig the belief system with the appropriate amount of moral certainty and superiority over outsiders, so as to provide all the pleasures of moral condemnation. In other words, you'd fill it with moral outrage porn. Third, if it were available, you'd want to entrench that belief system in a communication platform that awarded its users plenty of clear, direct affirmation for agreeing with each other. For that,

the gamified setting of Twitter will do quite nicely. Echo chambers instrumentalize our trust; moral outrage porn instrumentalizes our morality; and gamification instrumentalizes our goals.[33]

Bibliography

Schull, Natasha Dow. 2012. *Addiction by Design*. Princeton: Princeton University Press.

Bogost, Ian. 2011. "'Gamification Is Bullshit.'" The Atlantic. August 9, 2011. https://www.theatlantic.com/technology/archive/2011/08/gamification-is-bullshit/243338/.

Chou, Yu-kai. 2015. *Actionable Gamification: Beyond Points, Badges and Leaderboards*. Fremont, CA: CreateSpace Independent Publishing Platform.

Espeland, Wendy Nelson, and Michael Sauder. 2016. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. New York, New York: Russell Sage Foundation.

Frankfurt, Harry G. 2005. *On Bullshit*. Princeton, NJ: Princeton University Press.

Fricker, Elizabeth. 2006. "Second-Hand Knowledge." *Philosophy and Phenomenological Research* 73 (3): 592–618.

Frost-Arnold, Karen. 2014. "Trustworthiness and Truth: The Epistemic Pitfalls of Internet Accountability." *Episteme* 11 (1): 63–81.

Gabrielle, Vincent. 2018. "How Employers Have Gamified Work for Maximum Profit." *Aeon Magazine*, October 10, 2018. https://aeon.co/essays/how-employers-have-gamified-work-for-maximum-profit.

Hong, Lu and Scott Page. 2001. "Problem Solving by Heterogenous Agents." *Journal of Economic Theory* 97 (1) 123-63.

---. 2004. "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers." *Proceedings of the National Acdaemy of Sciences of the United States* 101 (46): 16385-89.

Huizinga, Johan. 1971. *Homo Ludens: A Study of the Play-Element in Culture*. Reprint edition. Beacon Press.

Jamieson, Kathleen Hall, and Joseph Cappella. 2010. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford: Oxford University Press.

Kretchmar, Scott. 2012. "Competition, Redemption, and Hope." *Journal of the Philosophy of Sport* 39 (1): 101–16.

Landemore, Hélène. 2013. *Democratic Reason*. Princeton: Princeton University Press.

Lugones, Maria. 1987. "Playfulness, 'World'-travelling, and Loving Attention."*Hypatia* 2 (2): 3-19.

[33] I'd like to thank Mark Alfano, Matthew Carlson, Helen Daly, Jon Ellis, Max Hayward, Aaron James, Jennifer Lackey, Michael Lynch, Elijah Millgram, Alison Rieheld, Adriel Trott, and Matt Strohl for their help with this paper. Key ideas for this paper emerged from my work with Bekka Williams moral outrage porn — including the notion of instrumentalization.

Lupton, Deborah, and Gavin JD Smith. 2017. "'A Much Better Person': The Agential Capacities of Self-Tracking Practices." SSRN Scholarly Paper ID 3085751. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3085751.

Madrigal, Alexis C. 2013. "The Machine Zone: This Is Where You Go When You Just Can't Stop Looking at Pictures on Facebook." The Atlantic. July 31, 2013. https://www.theatlantic.com/technology/archive/2013/07/the-machine-zone-this-is-where-you-go-when-you-just-cant-stop-looking-at-pictures-on-facebook/278185/.

McGonigal, Jane. 2011. *Reality Is Broken: Why Games Make Us Better and How They Can Change the World*. New York: Penguin Books.

McLuhan, Marshall. 1964. *Understanding Media: The Extensions of Man*. Signet Books.

Medina, Jose. 2012. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and Resistant Imaginations*. Oxford, New York: Oxford University Press.

Merry, Sally Engle. 2016. *The Seductions of Quantification: Measuring Human Rights, Gender Violence, and Sex Trafficking*. Chicago: University of Chicago Press.

Miller, Boaz, and Isaac Record. 2013. "Justified Belief in a Digital Age: On the Epistemic Implications of Secret Internet Technologies." *Episteme* 10 (2): 117–134.

National Public Radio. 2014. "Stuck in The Machine Zone: Your Sweet Tooth For 'Candy Crush.'" *NPR.Org*, 2014. https://www.npr.org/sections/alltechconsidered/2014/06/07/319560646/stuck-in-the-machine-zone-your-sweet-tooth-for-candy-crush.

Nguyen, C. Thi. 2017. "Competition as Cooperation." *Journal of the Philosophy of Sport* 44 (1): 123–137.

———. 2018. "Echo Chambers and Epistemic Bubbles." *Episteme*, 1–21. https://doi.org/10.1017/epi.2018.32.

———. 2019. "Games and the Art of Agency." *Philosophical Review* 128 (4): 423-462.

———. 2020. *Games: Agency as Art*. New York: Oxford University Press.

---------. Forthcoming. "Trust and Sincerity in Art". *Ergo*.

Nguyen, C. Thi, and Matthew Strohl. 2019. "Cultural Appropriation and the Intimacy of Groups." *Philosophical Studies* 176 (4): 981–1002. https://doi.org/10.1007/s11098-018-1223-3.

Nguyen, C. Thi, and Bekka Williams. 2020. "Moral Outrage Porn." *Journal of Ethics and Social Philosophy*.

Perrow, Charles. 2014. *Complex Organizations: A Critical Essay*. Brattleboro, Vermont: Echo Point Books & Media.

Porter, Theodore. 1996. *Trust in Numbers*. Princeton: Princeton University Press. https://press.princeton.edu/titles/5653.html.

Rini, Regina. 2017. "Fake News and Partisan Epistemology." *Kennedy Institute of Ethics Journal* 27 (S2): 43–64. https://doi.org/10.1353/ken.2017.0025.

Schull, Natasha Dow. 2012. *Addiction by Design*. Princeton: Princeton University Press.

Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven;: Yale University Press.

Seymour, Richard. 2019. "The Machine Always Wins: What Drives Our Addiction to Social Media." *The Guardian*, August 23, 2019, sec. Technology. https://www.theguardian.com/technology/2019/aug/23/social-media-addiction-gambling.

Stenros, Jaakko. 2012. "In Defence of a Magic Circle: The Social and Mental Boundaries of Play." http://www.digra.org/wp-content/uploads/digital-library/12168.43543.pdf.

Strohl, Matt. 2017. “Against Rotten Tomatoes.” *Aesthetics for* Birsd. September 21, 2017. https://aestheticsforbirds.com/2017/09/21/against-rotten-tomatoes/.

Suits, Bernard, and Thomas Hurka. 2014. *The Grasshopper - Third Edition: Games, Life and Utopia*. Peterborough, Ontario: Broadview Press.

Sunstein, Cass R. 2009. *Republic.Com 2.0*. Princeton, N.J.: Princeton University Press.

Tufekci, Zeynep. 2017. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. New Haven, CT, USA: Yale University Press.

———. 2018. “It’s the (Democracy-Poisoning) Golden Age of Free Speech.” *Wired*. https://www.wired.com/story/free-speech-issue-tech-turmoil-new-censorship/.

Velleman, David. 2000. “On the Aim of Belief.” In *The Possibility of Practical Reason*, 244–81. Oxford: Oxford University Press.

Waern, Annika. 2012. “Framing Games.” *DiGRA Nordic ’12: Proceedings of 2012 International DiGRA Nordic Conference* 10. http://www.digra.org/wp-content/uploads/digital-library/12168.20295.pdf.

Wedgwood, Ralph. 2002. “The Aim of Belief.” *Philosophical Perspectives* 16: 267–97.

Weimer, Steven. 2012. “Consent and Right Action in Sport.” *Journal of the Philosophy of Sport* 39 (1): 11–31.

Williams, Bernard. 1970. “Deciding to Believe.” In *Problems of the Self*, 136–51. Cambridge University Press.